

Evaluating Total Portfolio Performance: Pitfalls and Solutions

AVTORJI: Laurent Barras, Alexander D. Beath, Sebastien Betermier

Povzetek: Raziskava obravnava, kako se izvaja ocena celotne portfeljne performansne, z uporabo referenčnega portfelja kot odkaznega modela. Raziskovalci proučijo, kako izboljšanje portfelja v zvezi s širšo diverzifikacijo, boljšo strategično alokacijo kapitala in aktivno menovanje vpliva na merjenje vrednostnega dodatka. Natančno ugotovljeno je, da za učinkovito razlikovanje med vrednostnega dodatka in šumom potrebno je več kot 20 let performansnega podatka, kar je znatno več, kot je običajno uporabljeno v praksi. Raziskovalci ponudijo metode za merjenje in interpretacijo nesigurnosti, razdelitev vrednostnega dodatka in izboljšanje izbora odkaznega modela.

Metoda: Raziskovalci uporabljajo standardno merilo srednje variјancne optimizacije, da izdelajo optimalan portfelj, ki se odliča od referenčnega portfelja z treh različnih kanalov: širše diverzifikacije, boljše strategične alokacije kapitala in aktivne menovanje. Natančno ugotovljeno je, da 20 let performansnega podatka potrebujeta za učinkovito razlikovanje med vrednostnega dodatka in šuma.

KLJUČNE UGOTOVITVE:

- Referenčni portfelj test je zelo šumljiv, če je portfelj odličen od referenčnega.
- Več kot 20 let performansnega podatka potrebujeta za učinkovito razlikovanje med vrednostnega dodatka in šuma.
- Nekaj 36-37% šans je, da se portfelj ne uspešno zgodovi nad referenčnim portfeljem v eno leto, in 13-15% v deset let.

POMEN ZA TRGOVANJE: Raziskava podpira uporabo večjega časa za oceno performans, saj je 20 let potrebno za učinkovito razlikovanje med vrednostnega dodatka in šuma. To pomeni, da se morajo menjava in osebne sredstva izmenjati z večjim odborom in določenjem, da je 20 let dovolj za učinkovito oceno.

EVALUATING TOTAL PORTFOLIO PERFORMANCE: PITFALLS AND SOLUTIONS^{*}

Laurent Barras, Alexander D. Beath, Sebastien Betermier

September 9, 2026

Abstract

Reference portfolios of listed equities and government bonds are increasingly used as benchmarks to evaluate the investment skill of institutional investors at the total portfolio level. We show that portfolio optimization away from the reference portfolio creates an inherent tension in this approach: the more a fund exploits sources of value-add, the more its portfolio deviates from the reference portfolio, generating tracking error that makes benchmark-relative performance a noisier measure of skill. Using a stylized fund calibrated to realistic levels of diversification, capital market assumptions, and manager and security selection, we find that more than 20 years of performance data are required to reliably distinguish skill from noise, far longer than the three- to five-year horizons typically used in practice. We propose approaches to quantify and interpret this uncertainty, isolate sources of value creation, and improve benchmark selection.

Keywords: Performance evaluation, capital allocation, governance, institutional investors, total portfolio approach

JEL Classification: G11, G23

^{*}Barras is a Full Professor in Asset Management at University of Luxembourg; laurent.barras@uni.lu. Beath is Founder and CEO of Alex Beath & Associates; alex@alexbeath.com. Betermier is an Associate Professor of Finance at the Desautels Faculty of Management, McGill University, and the Executive Director of the International Centre for Pension Management (ICPM); sebastien.betermier@mcgill.ca. We thank Kevin Chang, Adrian Trollor, and participants at an ICPM Working Group session and a CPP Investments seminar for their insightful comments. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the organizations with which the authors are affiliated. Any errors are the authors' own.

Institutional asset owners such as pension, insurance, endowment and sovereign wealth funds make a wide range of investment decisions. They determine the appropriate level of risk exposure, define the long-term strategic asset allocation (SAA), and select individual funds and securities. Given the substantial size of these funds and their broad investment mandates, it is essential for fund managers, board members, and policy-makers to ensure that investment strategies are robust and, particularly in the case of public-sector funds, that public resources are used effectively. Accordingly, developing rigorous methods to evaluate overall portfolio performance is of critical importance.

The standard approach to evaluating investment performance is to define a benchmark portfolio and then compare the return generated by the fund’s management team to the benchmark return over a specified period of time (Wermers et al., 2026), usually between three and five years. This performance test is widely used in the world of active management because it provides a straightforward, transparent, and easy-to-interpret way to evaluate the value-add of active strategies over low-cost passive strategies. The return spread between a fund’s portfolio and the benchmark captures the value added — or lost — through the fund’s investment decisions.

Institutional investors have traditionally relied on this approach and constructed policy benchmarks as weighted combinations of asset-class benchmark returns, with weights determined by the funds’ SAA. However, because the policy benchmark is itself constructed from a fund’s SAA, excess returns relative to the benchmark do not capture the value created by the fund’s SAA decision. This is particularly limiting given that asset allocation policy explains over 90% of the variability in returns of a typical fund over time (Brinson et al., 1986, 1991; Ibbotson and Kaplan, 2000). Instead, excess returns measure only the value created through implementation decisions such as fund manager and security selection. As a result, traditional performance evaluation does not capture all dimensions of investment skill.

This limitation has motivated the search for alternative benchmarks that evaluate a fund’s overall investment strategy. One benchmark that has gained prominence in recent years is the reference portfolio, which consists of low-cost investments in publicly traded

equities and bonds calibrated to match the fund’s risk appetite. Unlike a traditional policy benchmark, the reference portfolio is not constructed using the fund’s chosen SAA; instead, it represents how the fund’s assets could be passively invested in the absence of the fund. The return spread between the fund’s actual portfolio and the reference portfolio is therefore intended to measure the value created by the fund’s complete investment strategy.

In this paper, we investigate how informative the reference portfolio is as a benchmark for evaluating investment skill and whether institutional funds should consider alternative benchmarks. Using a standard mean-variance framework, we construct the optimal portfolio of a stylized fund that adds value relative to the reference portfolio through three distinct channels: (1) broader diversification across asset classes, (2) superior strategic asset allocation resulting from more accurate capital market assumptions (CMAs), and (3) active management through fund manager and security selection. We refer to these skill sets as diversification, strategic, and active skills, respectively. We then study the resulting performance test and assess its ability to correctly distinguish value-creating from non-value-creating funds over practical evaluation horizons.

Specifically, we examine the frequency of two types of errors: (1) false negatives, in which the test fails to identify a fund that has superior skill and (2) false positives, in which the test concludes that a fund is skilled when the fund is not. We first quantify these error probabilities in a calibrated institutional setting and examine how they vary with different sources of value-add and tracking error. We then develop a theoretical model based on the mean-variance framework to identify benchmark specifications that minimize the false-negative probability for a skilled fund.

Our central result is that the reference portfolio test is highly noisy. Even over 10-year evaluation horizons, the test produces high rates of false negatives and false positives, frequently leading to incorrect conclusions about the skill of a fund’s management team. The key issue is that the informativeness of the reference portfolio as a benchmark depends on how closely the fund’s portfolio aligns with it. Yet portfolio optimization away from the reference portfolio increases tracking error, reducing the informativeness of the

performance test. This problem is compounded by uncertainty about expected returns, which leads to further differences between the optimized and reference portfolios and, consequently, additional tracking error.

To illustrate this limitation, we calibrate the mean-variance framework to a realistic institutional setting using forecasted asset returns and covariances from several CMA providers, together with active management performance estimates from CEM Benchmarking, a global benchmarking company. The fund's optimized portfolio delivers a meaningful value-add of 1.2-1.4% per year relative to the reference portfolio net of costs while remaining closely aligned with it, as evidenced by a return correlation above 90% and a volatility differential below 1%. Yet, even these values do not provide sufficient alignment for the reference portfolio test to reliably identify skill. There is a 36-37% chance that the fund underperforms the reference portfolio over a one-year horizon, and a 13-15% chance over a 10-year horizon. It takes more than 20 years for the false-negative probability to fall below 5% - a threshold that is widely used as the conventional significance level in scientific research. Thus, the performance test has a high false-negative rate, particularly over shorter horizons.

We next consider the opposite error: false positives. When the fund has zero value-add, we find that the reference portfolio test can still produce seemingly strong evidence of skill. Even when the realized average annual outperformance is 1% over a 10-year period, there is a greater than 20% probability of observing such outperformance purely by chance. This result highlights the difficulty of ruling out luck even when realized performance is strong. Together, these findings demonstrate the limited ability of reference portfolio tests to distinguish skill from noise over realistic investment horizons.

To understand how noise in performance measurement should inform benchmark design, we develop a theoretical model based on the mean-variance framework. The model distills the total portfolio problem to its essential features, allowing us to characterize the noise associated with each source of value-add and to vary the magnitude of each source in a tractable and controlled way. We use the model to identify the conditions under which alternative benchmarks are more effective than the reference portfolio at detecting

investment skill.

The model identifies a novel source of tracking error arising directly from portfolio optimization. Value-add generated through diversification and strategic portfolio optimization requires successive re-optimizations, each of which changes the portfolio weights relative to the reference portfolio and thereby generates additional tracking error. This source of tracking error is distinct from that associated with active management, which is implemented within the fund's SAA and therefore does not require re-optimizing the portfolio.

A second channel through which portfolio optimization generates tracking error arises from differences in return forecasts. When the fund's CMAs are weakly correlated with those of individual investors and other funds, even modest differences in forecasts can lead to substantially different optimized portfolios and, consequently, greater tracking error. By contrast, tracking error from active management is independent of forecast uncertainty because active management is implemented within the fund's SAA.

Taken together, these two channels create a trade-off in benchmark selection. Including multiple sources of value-add increases the signal available to detect investment skill, but portfolio re-optimization also generates tracking-error noise that can offset this benefit. A total portfolio test is therefore more informative only when the fund's overall value-add is sufficiently large to outweigh this noise. For example, in our calibrated setting, the fund's CMAs are strongly correlated with those of individual investors, while active-management value-add and tracking error are 0.29% and 1.9%, respectively. We find that the total portfolio test has a lower false-negative risk than a test of only active management when the combined value-add from asset class diversification and strategic portfolio optimization exceeds 0.40%. This threshold rises to 1.25% when the fund's CMAs are weakly correlated with those of individual investors.

When tracking-error noise is high, performance tests that exclude noisier sources of value-add may provide a more precise measure of investment skill. In particular, we find that using a basic portfolio diversified across all asset classes as the benchmark, thereby embedding diversification skill in the benchmark and excluding it from the performance

test, results in a lower false-negative risk than using the reference portfolio.

From these findings, we draw several conclusions about benchmarking total fund performance. Separating skill from luck at the total portfolio level is extremely difficult, and performance tests based on a fund’s total portfolio should therefore be interpreted with caution. Noise in total portfolio performance also makes it difficult to identify the sources of investment skill. The false-positive and false-negative probabilities provide useful tools to quantify and interpret this uncertainty.

Our findings also highlight an inherent tension between using the reference portfolio as a performance benchmark and optimizing the total portfolio away from it, as in the total portfolio approach (TPA). TPA allows a fund’s management team to leverage its investment expertise and optimize the portfolio (Elkamhi and Lee, 2025). However, as the portfolio deviates further from the reference portfolio, benchmark-relative performance becomes a noisier measure of skill.

Finally, the high tracking error associated with diversification across asset classes raises important questions about benchmark selection and performance evaluation. If the value of diversification is well established but its associated tracking error is large, should funds exclude this source of value creation from the performance evaluation exercise, as our theoretical results suggest? More broadly, should funds reconsider how performance is evaluated when portfolio optimization itself generates substantial tracking error? One complementary approach would be to assess the quality of investment processes *ex ante*, rather than rely solely on outcomes *ex post*. For example, independent external reviews could assess whether funds maintain efficient cost structures, robust governance processes, and coherent investment strategies. Such reviews would provide additional insights into the quality of a fund’s investment process.¹

Our paper contributes to several strands of the literature. First, relative to the literature on performance evaluation, which primarily focuses on evaluating active investment strategies (Grinblatt and Titman, 1995; Grinold and Kahn, 2000), we evaluate the perfor-

¹Shen et al. (2026) argue that, under TPA, benchmark comparisons are no longer sufficient on their own.

mance of a fund’s total portfolio, which incorporates three distinct sources of value-add. We show that portfolio optimization generates a new component of tracking error, separate from the tracking error associated with traditional active management. Second, relative to the literature on portfolio choice with heterogeneous beliefs (Back, 2017; Basak, 2000), we develop a tractable specification of the distortion in investors’ return forecasts and show how it generates another source of tracking error. Finally, relative to the literature on benchmark design for institutional investors (Blake and Timmermann, 2002), we provide a tractable framework for assessing how different combinations of value-add and tracking error affect the ability of alternative benchmarks to identify investment skill in a fund’s total portfolio.

The rest of the paper is organized as follows. Section 1 provides a brief historical review of the growing use of the reference portfolio as a performance management and governance tool. Section 2 presents the mean-variance framework for a fund that generates multiple sources of value-add, reviews the performance test used to evaluate managerial skill, and introduces false-negative and false-positive probabilities as measures of test informativeness. Section 3 calibrates the framework to a realistic institutional setting and quantifies the false-negative and false-positive probabilities. Section 4 develops the theoretical model and examines optimal benchmark design. Section 5 discusses implications for practice and policy, and Section 6 concludes. Details about the data and model derivation are provided in the Appendix.

1 Brief History of the Reference Portfolio

The use of the reference portfolio as a performance management and governance tool has increased substantially in recent years among large institutional investors such as pension funds, sovereign wealth funds, and endowments. Initially implemented by CPP Investments and the New Zealand Superannuation Fund in the mid-2000s,² the reference portfolio framework is now widely used by institutional investors across multiple countries

²See Raymond (2009) for an explanation of CPP Investments’ original approach.

such as GIC in Singapore, CalPERS in the United States, and NSW Treasury Corporation in Australia.³ The specific use of the reference portfolio varies across funds, ranging from a tool to communicate and set risk appetite at the board to a formal benchmark for performance evaluation and incentive compensation.

The growing use of the reference portfolio reflects the expanding responsibilities entrusted to fund management teams under a “Total Portfolio Approach.” In this framework, investment decisions are made at the total portfolio level rather than within isolated asset-class mandates. Management teams are granted greater discretion over tactical allocation, implementation, liquidity management, leverage, and cross-asset diversification decisions. The reference portfolio serves as a natural benchmark against which the value-add of these broader discretionary decisions can be evaluated.

Beyond its use by the funds themselves, the reference portfolio has become an attractive benchmark for the media. This is illustrated by an editorial published by the Chicago Tribune on August 5, 2025,⁴ titled, “Chicago’s pension investments likely are doing worse than your 401(k). Big problem.” Because the reference portfolio represents how the assets might be invested in the absence of the fund, it provides a natural benchmark to assess the fund’s complete investment strategy, relative to what investors could achieve by investing on their own.

2 Assessing the Reference Portfolio as a Performance Benchmark

To assess the effectiveness of the reference portfolio as a benchmark for detecting skill, we develop a standard mean-variance framework for an institutional fund that generates value through multiple sources. This framework provides the foundation for the subsequent calibration exercise and theoretical model. In this section, we lay out the conceptual framework, define the key portfolios used throughout the analysis, review

³See <https://i3-invest.com/2026/07/reference-portfolios-what-are-funds-really-benchmarking/>

⁴<https://www.chicagotribune.com/2025/08/20/editorial-chicago-pensions-investment-performance/>

the performance test used to evaluate managerial skill, and introduce false-negative and false-positive probabilities as measures of test informativeness.

2.1 Mean-Variance Framework and Key Portfolios

The investment universe comprises N asset classes such as listed equities, bonds, and private real estate. Each asset class has a representative benchmark fund with a stochastic return. The benchmark fund could be either a passively managed index fund or another publicly available investment vehicle. Investors do not observe the true expected returns, volatilities, or correlation matrix of the benchmark funds. Instead, they form forecasts, which we refer to collectively as a set of capital market assumptions (CMA).

Individual investors seek to maximize the expected return of their portfolio without exceeding volatility $\bar{\sigma}$. If they invest on their own, investors only have access to the benchmark funds of listed stocks and bonds. The *reference portfolio* R , which has expected return μ_R and volatility σ_R , represents the optimal portfolio of benchmark funds in listed equities and bonds based on a CMA that is publicly available to individual investors. Thus, the reference portfolio captures how individuals would passively invest on their own.

Alternatively, investors can put their savings into an institutional fund that creates value in three distinct ways. The fund is able to invest across the N asset classes and therefore benefits from enhanced diversification. The fund also has a superior CMA and thus optimizes the portfolio under a more accurate set of forecasts. Finally, the fund actively manages assets and generates higher returns on average than the benchmark funds. We refer to these three skill sets as diversification, strategic, and active skills, respectively. The fund's *actual portfolio* has expected return μ_A and volatility σ_A . Its value-add relative to what investors can do on their own is therefore $\alpha = \mu_A - \mu_R$. We provide further details on the optimization of the fund's actual portfolio below.

To isolate each skill set, we define two additional portfolios. The *diversified portfolio* D , which has expected return μ_D and volatility σ_D , represents the optimal portfolio of benchmark funds in the $N > 2$ asset classes, calibrated to have the same volatility of $\bar{\sigma}$

and based on a CMA of comparable quality to that publicly available to investors. The average return spread between the diversified and reference portfolios, $\alpha_D = \mu_D - \mu_R$, isolates the value-add resulting from diversification across the N asset classes. We refer to α_D as diversification value-add.

The *strategic portfolio* S , which has expected return μ_S and volatility σ_S , represents the optimal portfolio of the N benchmark funds, calibrated to have volatility $\bar{\sigma}$ and based on the fund's superior CMA. This portfolio reflects the fund's SAA decision. The average return spread between the strategic and diversified portfolios, $\alpha_S = \mu_S - \mu_D$, isolates the value-add resulting from the superior optimization of portfolio weights across the N asset classes. We refer to α_S as strategic value-add.

In line with common practice in the industry, we assume that active management is implemented within the fund's SAA and therefore does not require re-optimizing the portfolio. In other words, the fund's actual portfolio has the same weights as the strategic portfolio, but different returns that reflect its active investments. The actual portfolio has a tracking error of $\sigma_{\epsilon,A}$ relative to the strategic portfolio. The average return spread between the actual and strategic portfolios, $\alpha_A = \mu_A - \mu_S$, isolates the value-add from active management. We refer to α_A as active value-add.

The four portfolios and their distinct properties are summarized in Figure 1. The fund's total value-add is the sum of each value-add:

$$\alpha = \mu_A - \mu_R = \alpha_D + \alpha_S + \alpha_A. \quad (1)$$

The average return spread between the active and reference portfolios captures the total value created through the fund's enhanced diversification, superior strategic asset allocation, and active management.

Each source of value-add can be assessed individually by comparing adjacent portfolios. For example, the active value-add, α_A , can be evaluated by comparing the expected returns of the actual and strategic portfolios. It is also possible to evaluate combinations of value-add sources. For example, the sum of active and strategic value-adds, $\alpha_A + \alpha_S$,

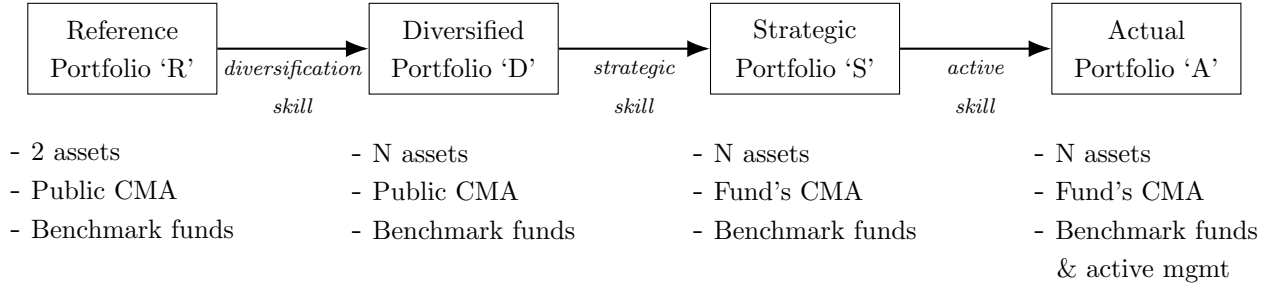


Figure 1: **Properties of the four portfolios** This figure summarizes the key properties of the reference, diversified, strategic, and actual portfolios in terms of the number of assets, capital market assumptions, and types of investment funds.

can be assessed by comparing the expected returns of the actual and diversified portfolios. Evaluating total portfolio performance requires assessing all sources of value jointly by comparing the expected returns of the actual and reference portfolios.

2.2 Performance Evaluation

Because the expected returns of the different portfolios are unobservable, performance evaluation relies on comparing realized average returns across portfolios over a given time period. This comparison, however, can confound investment skill with luck, potentially leading to two types of errors: failing to recognize skill when poor realized performance is driven by bad luck (a false negative), or attributing strong realized performance to skill when it is instead driven by good luck (a false positive). In this section, we review the traditional performance test and introduce two metrics — the false-negative and false-positive probabilities — to quantify the extent to which realized performance can lead to these errors.

2.2.1 The Performance Test

We consider a textbook setting where the fund's actual portfolio A and the reference portfolio R (the benchmark) have realized returns $r_{A,t}$ and $r_{R,t}$ at time $t \in \{1, \dots, T\}$ that

are given by the following processes:

$$r_{A,t} = \mu_A + \epsilon_{A,t}, \quad (2)$$

$$r_{R,t} = \mu_R + \epsilon_{R,t}, \quad (3)$$

where the shocks $\epsilon_{A,t}$ and $\epsilon_{R,t}$ are i.i.d. across time and jointly normally distributed with zero means, standard deviations σ_A and σ_R , and correlation ρ_{AR} .⁵ The fund has superior investment skill if $\alpha = \mu_A - \mu_R > 0$. The mean parameters μ_A and μ_R and the corresponding value-add α are not observed.

The common industry practice is to assess fund skill by comparing the average realized returns of both portfolios over the sample period, $\bar{r}_A = \frac{1}{T} \sum_t (\mu_A + \epsilon_{A,t})$ and $\bar{r}_R = \frac{1}{T} \sum_t (\mu_R + \epsilon_{R,t})$.⁶ We denote the average return spread in the sample by

$$\bar{\alpha} = \bar{r}_A - \bar{r}_R = \alpha + \frac{1}{T} \sum_t (\epsilon_{A,t} - \epsilon_{R,t}), \quad (4)$$

and its realized value by $\bar{\alpha}_{obs}$. Under this performance test, a fund is deemed skilled if $\bar{\alpha}_{obs} > 0$ and unskilled otherwise. This test can be applied to any pair of portfolios. We refer to the procedure above as the *A* vs. *R* (*AvR*) test and use analogous notation for subsequent portfolio comparisons.

As Equation (4) shows, the performance test is a noisy indicator of investment skill. The observed return spread, $\bar{\alpha}_{obs}$, can be positive either because the fund has superior investment skill or because it is lucky relative to the reference portfolio. The noise level

⁵This i.i.d. specification provides a baseline case in which asset returns are well behaved. In practice, the actual portfolio may exhibit additional features that depart from this baseline, including volatility smoothing and serial correlation because private assets are valued infrequently and reported returns are affected by stale prices. These features, in turn, may increase tracking error relative to the reference portfolio through differences in measured volatility and imperfect correlation between the portfolios. In addition, positive serial correlation may reduce the effective amount of independent information in the sample and therefore increase the noise in performance tests relative to the i.i.d. benchmark.

⁶Oftentimes, the comparison focuses on the cumulative returns of the two portfolios, which are not normally distributed. Here, we instead focus on arithmetic returns because they tend to be better behaved: over sufficiently long periods, their distribution is approximately normal, and the return spread between the two portfolios is also approximately normally distributed.

is captured by the tracking error, defined as the standard deviation of the return spread:

$$\sigma_\epsilon = \sqrt{Var[\epsilon_{A,t} - \epsilon_{R,t}]} = \sqrt{\sigma_A^2 + \sigma_R^2 - 2\rho_{AR}\sigma_A\sigma_R} \quad (5)$$

$$= \sqrt{(\sigma_A - \sigma_R)^2 + 2 \cdot \sigma_A\sigma_R(1 - \rho_{AR})}. \quad (6)$$

Tracking error has two key drivers: the imperfect correlation between the two return series and their volatility differential.

It is useful to assess the size of the return spread relative to the underlying noise. We therefore express the average return spread per unit of standard deviation:

$$\bar{z} = \sqrt{T} \frac{\bar{\alpha}}{\sigma_\epsilon}, \quad (7)$$

and denote its realized value by $\bar{z}_{obs}$. Under the assumptions above, $\bar{z}$ follows a normal distribution with mean

$$z = \sqrt{T} \frac{\alpha}{\sigma_\epsilon} \quad (8)$$

and variance equal to one.

To quantify the ability of the *AvR* test to distinguish skill from luck, we examine the probabilities of false negatives and false positives associated with different values of the realized average return spread, $\bar{\alpha}_{obs}$. Both types of errors are relevant, but in different situations and from different perspectives.

2.2.2 False Negatives

False negatives are particularly concerning for fund management when realized value-add is negative, corresponding to underperformance relative to the benchmark. In this setting, a skilled fund may appear unskilled because unfavorable realized performance obscures its underlying ability. External stakeholders may therefore question whether the fund continues to generate sufficient value. Such a misclassification can have significant consequences, including lower incentive compensation, staff reductions or terminations,

and strategy recall.

To quantify the risk of a false-negative outcome, we define the false-negative probability as the probability that a skilled fund produces a realized performance spread below the observed value:

$$P(T) = P[\bar{\alpha} < \bar{\alpha}_{obs} \mid \alpha > 0] = \Phi(\bar{z}_{obs}; z, 1), \quad (9)$$

where T denotes the sample length and $\Phi(\bar{z}_{obs}; z, 1)$ denotes the cumulative distribution function of a normal distribution with mean z and variance one, evaluated at $\bar{z}_{obs}$.

The false-negative probability $P(T)$ is low when $\bar{z}_{obs}$ is low relative to the true standardized value-add, z . As Equation (8) shows, the latter increases with the sample length T and with the test's signal-to-noise ratio, α/σ_ϵ . A high signal-to-noise ratio corresponds to a setting in which the fund's actual portfolio has high value-add and low tracking error relative to the reference portfolio. Importantly, $P(T)$ does not depend on the sampling frequency. As Merton (1980) explains, increasing the sampling frequency does not provide additional information about the average return when the total calendar time is fixed.

The false-negative probability provides a guide for interpreting negative realized performance. A high value of $P(T)$ indicates that unfavorable performance is relatively likely even for a skilled fund and should therefore be interpreted with caution. In contrast, a low value of $P(T)$ indicates that such underperformance would be unlikely for a skilled fund, providing greater confidence that it instead reflects a lack of skill.

2.2.3 False Positives

False positives are particularly concerning for external observers when realized value-add is positive, corresponding to outperformance relative to the benchmark. In this setting, fund management may view favorable realized performance as evidence of superior skill, even though the performance instead reflects good luck. An external observer who does not know the fund's true level of skill may therefore question whether the observed

outperformance reflects superior skill or good luck. Such a misclassification may lead to greater capital allocation to the fund and higher compensation for the management team.

To quantify the risk of incorrectly attributing luck to skill, we define the false-positive probability as the probability that a fund with no superior skill produces a realized performance spread greater than the observed value:

$$P_0(T) = P[\bar{\alpha} > \bar{\alpha}_{obs} \mid \alpha = 0] = 1 - \Phi(\bar{z}_{obs}; 0, 1). \quad (10)$$

This probability is high when $\bar{z}_{obs}$ is close to zero. Because $\bar{z}_{obs}$ expresses the observed performance spread relative to its noise, a higher tracking error, σ_ϵ makes skill more difficult to identify.

The false-positive probability provides a guide for interpreting positive realized performance. A high value of $P_0(T)$ indicates that favorable performance is relatively likely even without superior skill and should therefore be interpreted with caution. In contrast, a low value of $P_0(T)$ indicates that such outperformance is unlikely without superior skill, providing greater confidence that it reflects skill rather than luck.

2.3 Benchmark Selection

The false-negative and false-positive probabilities measure the reliability of the performance test. In addition, the false-negative probability provides a natural criterion for benchmark selection: a fund can choose a benchmark ex ante to reduce the risk that its investment skill goes undetected. We consider a threshold of $\bar{\alpha}_{obs} = 0$, which corresponds to performance in line with the benchmark, and ask which benchmark minimizes the false-negative probability $P(T)$. Specifically, is it more informative to use the reference portfolio as a benchmark that captures all sources of skill, or to benchmark individual sources of skill separately?

As a first comparison, we examine the total portfolio test (AvR) and the active skill test, which compares the actual portfolio with the strategic portfolio (AvS). Given

Equation (8), the *AvR* test produces fewer false negatives than the *AvS* test if

$$\frac{\alpha}{\sigma_{\epsilon}} > \frac{\alpha_A}{\sigma_{\epsilon,A}}, \quad (11)$$

and more false negatives otherwise.⁷ Said differently, it is preferable to combine the different sources of investment skill into a single performance test if

$$\frac{\alpha}{\alpha_A} > \frac{\sigma_{\epsilon}}{\sigma_{\epsilon,A}}. \quad (12)$$

The *AvR* test is more informative about investment skill than the *AvS* test when the proportional increase in signal from additional sources of value-add exceeds the proportional increase in tracking error, and less informative when the opposite is true. An additional limitation of the combined test is that it does not identify the sources of value-add underlying total portfolio performance.

In the sections that follow, we quantify the likelihood of false negatives and false positives by applying the performance test to the fund’s total portfolio and to multiple portfolios along the value chain. We use these probabilities to assess (1) the reliability of the performance test and (2) how benchmark selection can reduce the risk of false negatives. We examine each issue separately.

3 Distinguishing Skill From Luck: A Calibration Exercise

3.1 Calibrating the Mean-Variance Framework

To calibrate our framework and apply the performance tests, we proceed in three steps. First, we define the asset classes in which institutional funds invest. Second, we optimize the reference, diversified, strategic, and actual portfolios for a stylized fund that adds value along all three dimensions. To do so, we use forecasts of expected returns, volatili-

⁷We assume that the performance tests are conducted on the same sample.

ties, and correlations across these asset classes from well-established CMA providers. We also impose constraints on portfolio weights to ensure that the optimized portfolios are realistic. Third, we determine the risk-return properties of each portfolio using a proxy for the true distribution of asset returns, along with estimates of active value-add from CEM Benchmarking. Given these inputs, we evaluate the portfolios using the performance tests and assess the tests’ ability to distinguish skill from luck.

Our approach offers several advantages. It provides a realistic setting for modeling the portfolio of an institutional fund with multiple sources of investment skill. It is consistent with recent work by Andonov et al. (2026), which shows that differences in US public pension funds’ portfolio allocations can be explained by mean-variance optimization under heterogeneous beliefs and portfolio constraints. Our approach also allows us to quantify the true magnitude of each source of value-add in a setting where the true return distribution is known. This, in turn, enables us to assess how frequently the performance test correctly identifies the portfolio with greater value-add and to compare the informativeness of different performance tests at different stages along the value chain.

Step 1: Asset Classes The first step is to define the asset classes in which institutional funds invest. We consider three main asset classes: equities, fixed income, and real estate, which together capture over 90% of assets in institutional portfolios (Beath and Flynn, 2023). We further divide each asset class into ownership status (i.e., public vs. private), portfolio company cap size (listed equities), region (listed equities) and duration (fixed income). This level of division yields nine asset classes: (i) US large cap equity, (ii) US small cap equity, (iii) non-US equity, (iv) private equity, (v) aggregate US bonds, (vi) long-duration bonds, (vii) private real estate, (viii) listed real estate, and (ix) cash.

Step 2: Optimization of Portfolios The second step is to optimize the portfolios based on return forecasts. We use the publicly available CMA from the Callan Institute as the set of forecasts available to individual investors and refer to it as the public CMA (Kloepfer et al., 2021). The data represent projected asset class performance as of 2021,

with an investment horizon of 10 years. The dataset provides the annualized arithmetic average returns and volatilities, along with a nine-by-nine correlation matrix. We assume that the return forecasts represent passive exposures to each asset class and are net of costs. For the fund’s CMA, we use a separate set of forecasts and assume that these forecasts coincide with the true distribution of returns, reflecting the fund’s strategic skill. We describe the fund’s CMA data in Step 3 below and provide additional details in the Appendix.

Given these inputs, we optimize each portfolio $i \in \{R, D, S, A\}$ by solving a constrained mean-variance optimization problem that maximizes expected return under a given CMA, subject to a volatility target of 10% and portfolio weight constraints. Short positions are not permitted. Allocations to listed equities and bonds must each be at least 10%, while allocations to other asset classes may not exceed 15% of total portfolio value. We impose these portfolio constraints to mimic the investment decisions of institutional investors and to avoid corner solutions in the mean-variance optimization.

The reference portfolio is the optimal portfolio of large-cap US equities and aggregate bonds based on the public CMA and the volatility target $\bar{\sigma} = 10\%$. The diversified portfolio is the optimal portfolio of the nine asset classes based on the public CMA and volatility target $\bar{\sigma} = 10\%$. The strategic portfolio is the optimal portfolio of the nine asset classes based on the fund’s CMA and volatility target $\bar{\sigma} = 10\%$. The actual portfolio has the same portfolio weights as the strategic portfolio, but its asset-class returns incorporate active value-add, which we discuss next.

Step 3: Risk-Return Properties of Each Portfolio The third step is to calculate the expected returns, volatilities, and correlations of each portfolio using a proxy for the true return distribution. We consider two proxies. The first uses the 2021 CMA from T. Rowe Price (Lustig et al., 2021), while the second uses historical annual net returns across asset classes for a sample of approximately 200 U.S. public- and corporate-sector defined-benefit (DB) pension funds over the period 1998–2021 from CEM Benchmarking

(Beath and Flynn, 2023).⁸

The two proxies have different strengths and limitations. Using another CMA provider captures the types of portfolio differences that can arise when institutional funds rely on different CMA providers (Andonov et al., 2026).⁹ However, CMAs may be correlated with one another and may differ substantially from the true return distribution. Using the realized return distribution provides a closer approximation to the true return distribution. However, it represents the sample distribution rather than the true underlying distribution. Taken together, the two approaches provide complementary perspectives and allow us to assess whether our results are robust to alternative proxies for the true return distribution.

The returns of the fund’s actual portfolio incorporate active returns. We obtain estimates of net value-add from the CEM Benchmarking database of institutional funds over the period 1992–2020 (Beath et al., 2022). We consider a fund with an active value-add of 0.29% and tracking error of 1.90%, corresponding to the average value-add (net of costs) and tracking error among funds with assets under management above USD 10 billion. We assume active returns are uncorrelated with the benchmark fund returns.

3.2 Portfolio Characteristics

Table 1 displays asset allocation weights for the four portfolios. The weights of the reference and diversified portfolios are both based on the public CMA. The reference portfolio is close to a “60/40” allocation.¹⁰ Since the Callan Institute CMA is unfavorable towards large US large cap equity allocations and favorable towards large private equity and real estate allocations, the diversified portfolio only holds a small proportion in listed

⁸The designation of the Callan CMA as the publicly available CMA and the T. Rowe Price CMA as the proxy for the true distribution is purely for illustrative purposes and does not imply any judgment regarding the relative accuracy of the providers’ forecasts.

⁹Several funds develop their CMAs internally. Our objective is not to claim that the two selected CMAs capture the full range of differences in forecasts across providers, but rather to illustrate the types of differences in forecasts that can arise among institutional investors.

¹⁰The 60-40 allocation partly reflects the 10% volatility target. A higher volatility target would allow the reference portfolio to hold a larger share of listed equities. One key benefit of reference portfolios for pension fund boards is that they provide a framework for calibrating the fund’s appropriate risk level.

equities (10%) and higher proportions in private equity (15%), private real estate (15%), and listed real estate (15%).

Table 1: **Asset Allocations.** The table reports asset allocation weights (in %) for the reference, diversified, strategic, and actual portfolios. Each portfolio is obtained by solving a constrained mean–variance optimization problem that maximizes expected return under a given set of capital market assumptions (CMA), subject to a 10% volatility target and the portfolio weight constraints reported at the bottom of the table. The reference and diversified portfolios are constructed using the 2021 CMA of the Callan Institute, while the strategic and actual portfolios are constructed using either the 2021 CMA of T. Rowe Price or historical returns for U.S. defined-benefit pension funds from CEM Benchmarking in 1998–2021. The reference portfolio comprises U.S. large-cap equities and aggregate bonds, whereas the diversified, strategic, and actual portfolios comprise all nine asset classes. The actual portfolio is identical to the strategic portfolio but its asset-class returns incorporate active value-add.

Portfolio	Large Cap US Stock	Small Cap US Stock	Non- US Stock	Private Equity	Agg- regate Bonds	Long Duration Bonds	Private Real Estate	Listed Real Estate	Cash
<i>True Return Distribution = Alternative CMA</i>									
Reference	56.67	0.00	0.00	0.00	43.33	0.00	0.00	0.00	0.00
Diversified	10.00	0.39	5.44	15.00	39.18	0.00	15.00	15.00	0.00
Strategic	10.00	7.44	15.00	14.07	38.48	0.00	15.00	0.00	0.00
Actual	10.00	7.44	15.00	14.07	38.48	0.00	15.00	0.00	0.00
<i>True Return Distribution = Historical Distribution</i>									
Reference	56.67	0.00	0.00	0.00	43.33	0.00	0.00	0.00	0.00
Diversified	10.00	0.39	5.44	15.00	39.18	0.00	15.00	15.00	0.00
Strategic	23.00	7.15	0.00	15.00	20.31	15.00	4.54	15.00	0.00
Actual	23.00	7.15	0.00	15.00	20.31	15.00	4.54	15.00	0.00
<i>Max:</i>	100.00	15.00	15.00	15.00	100.00	15.00	15.00	15.00	15.00
<i>Min:</i>	10.00	0.00	0.00	0.00	10.0	0.00	0.00	0.00	0.00

The weights of the strategic and actual portfolios depend on the fund’s CMA. When the alternative CMA is used, the portfolios maintain a small allocation to listed equities (10%) and large allocations to private equity (15%), non-US equities (15%), and private real estate (15%). This reflects the T. Rowe Price CMA’s relatively unfavorable outlook for large-cap US equities and listed real estate and favorable outlook for non-US stock, private equity, aggregate bonds, and private real estate. The resulting asset mix is relatively similar to the average asset allocation of US defined-benefit pension funds in 2021 (Beath and Flynn, 2023). In contrast, when the historical data is used, the strategic and actual portfolios shift allocations from non-US stock to large-cap US stock (23%), from

aggregate bonds to long duration bonds (15%), and from private real estate to listed real estate (15%), reflecting the strong realized performance of listed equities and listed real estate over the sample period.

Table 2 reports the average returns, volatilities, Sharpe ratios, and correlation structures of the four portfolios evaluated under each proxy for the true return distribution. When the alternative CMA is used, average returns increase from 3.91% for the reference portfolio to 5.13% for the actual portfolio, an increase of 1.22 percentage points. Volatility also increases, from 9.40% to 10.18%, indicating that part of the higher return reflects greater risk.¹¹ However, the Sharpe ratio also rises, from 0.42 to 0.50, indicating a material improvement in risk-adjusted performance. Each successive portfolio has a higher Sharpe ratio, reflecting superior performance at every step along the value chain. The diversification, strategic, and active value-adds are 0.73, 0.19, and 0.29, respectively. Despite the higher returns, the actual portfolio remains closely aligned with the reference portfolio. The volatility differential is small (0.78%), and the correlation between the two portfolios is high (93%).

The results are similar when the historical return distribution is used as proxy for the true distribution. Relative to the reference portfolio, the actual portfolio has an average return that is 1.40 percentage points higher, a Sharpe ratio that is 12 percentage points higher, a small volatility differential of 0.26, and a high correlation of 0.92. The diversification, strategic, and active value-adds are 0.25, 0.86, and 0.29, respectively. Taken together, these statistics indicate that a fund’s optimized portfolio provides economically significant value-add relative to the reference portfolio while remaining closely aligned in terms of risk.

3.3 False-Negative Probabilities

Suppose the fund has generated no excess return relative to its reference portfolio over the evaluation period, i.e. $\bar{\alpha}_{obs} = 0$. When external stakeholders question the fund’s

¹¹The true volatility of the reference and diversified portfolios is not necessarily 10%, since the CMA forecasts used to construct them differ from the true return distribution.

Table 2: **Portfolio Risk-Return Properties.** This table reports the average return, volatility, Sharpe ratio, and correlation structure of the reference, diversified, strategic, and actual portfolios. We assume that the true underlying return distribution corresponds to either the 2021 capital market assumptions from T. Rowe Price or the historical return distribution in 1998-2021 from CEM Benchmarking.

	Avg.		Sharpe	Correlations			
Portfolio	Return	Vol.	Ratio	Ref.	Div.	Strat.	Actual
<i>True Return Distribution = Alternative CMA</i>							
Reference	3.91	9.40	0.42	1.00	0.92	0.94	0.93
Diversified	4.64	9.95	0.47	0.92	1.00	0.98	0.96
Strategic	4.84	10.00	0.48	0.94	0.98	1.00	0.98
Actual	5.13	10.18	0.50	0.93	0.96	0.98	1.00
<i>True Return Distribution = Historical Distribution</i>							
Reference	8.25	9.92	0.83	1.00	0.87	0.94	0.92
Diversified	8.51	9.39	0.91	0.87	1.00	0.98	0.96
Strategic	9.36	10.00	0.94	0.94	0.98	1.00	0.98
Actual	9.65	10.18	0.95	0.92	0.96	0.98	1.00

performance, a natural question for fund management is: what is the likelihood that a genuinely skilled fund nonetheless performs no better than the reference portfolio over the evaluation period? We answer this question by calculating the false-negative probability of the performance test, first for the fund’s total portfolio and then by decomposing the result across multiple portfolios along the value chain.

For each pair of portfolios, we use the expected return spread from Table 2 (“true value-add”) and set the observed value-add, $\bar{\alpha}_{obs}$ to zero. Table 3 reports the true and observed value-adds, tracking error, and false-negative probability at one- and 10-year horizons. We also calculate T^* , the number of years required to reduce the false-negative probability below 5%, a widely used threshold in scientific research.¹²

Comparing the actual and reference portfolios, the results are striking. Despite the fund’s high value-add and close alignment with the reference portfolio, there is a 36-37% probability of underperforming the reference portfolio over a one-year horizon and a 13–15% probability over a 10-year horizon. It takes more than 20 years for the false-

¹²We obtain the corresponding value of $\bar{z}^* = 1.645$ for a one-sided test. The number of years required to reduce the false-negative probability below 5% is given by $T^* = \left(\frac{\bar{z}^* \sigma_\epsilon}{\alpha} \right)^2$.

Table 3: **False-Negative Probabilities.** This table reports the levels of value-add (true and observed), tracking error, and false-negative probabilities for a series of performance tests. Each test is based on the return spread between a portfolio and a selected benchmark. Value-add is measured as the average return spread, while tracking error is measured as the standard deviation of the return spread. The variable $P(T)$ denotes the probability that the realized value-add is below the observed value, conditional on the portfolio having positive true value-add, over either a one-year ($T = 1$) or 10-year ($T = 10$) investment horizon. The variable T^* denotes the number of years required for $P(T)$ to be equal to 5%. We assume that the underlying return distribution corresponds either to the 2021 capital market assumptions from T. Rowe Price or to the historical return distribution from 1998–2021 from CEM Benchmarking.

Test	Portfolio	Benchmark	Value-Add		Tracking Error	$P(T)$		T^*
			True	Obs.		1-year	10-years	
<i>True Return Distribution = Alternative CMA</i>								
DvR	Diversified	Reference	0.73	0.00	3.92	0.43	0.28	77.76
SvD	Strategic	Diversified	0.19	0.00	2.11	0.46	0.39	324.63
AvS	Actual	Strategic	0.29	0.00	1.90	0.44	0.31	116.86
AvR	Actual	Reference	1.22	0.00	3.78	0.37	0.15	26.35
<i>True Return Distribution = Historical Distribution</i>								
DvR	Diversified	Reference	0.25	0.00	4.98	0.48	0.44	1,041.63
SvD	Strategic	Diversified	0.86	0.00	2.16	0.35	0.11	17.40
AvS	Actual	Strategic	0.29	0.00	1.90	0.44	0.31	116.86
AvR	Actual	Reference	1.40	0.00	3.95	0.36	0.13	21.64

negative probability to fall below 5%. These results are similar for both proxies for the true return distribution. The performance test applied to the total portfolio therefore carries a meaningful risk of false negatives, particularly over shorter horizons.

We repeat the performance test for each pair of adjacent portfolios. In Panel A, the false-negative probabilities range from 43% to 46% over a one-year horizon and from 28% to 39% over a 10-year horizon, exceeding those for the total portfolio. In Panel B, the false-negative probabilities are particularly high for the diversification test (44% over a 10-year horizon) because value-add is low (0.25%) while tracking error is high (4.98%). Although tracking error is generally lower for each adjacent portfolio pair, value-add is also lower, making it more difficult to distinguish a lack of skill from bad luck. Thus, in this particular calibration the aggregate value-add across the fund makes the reference portfolio the most informative benchmark for detecting skill, although it remains a highly noisy benchmark in absolute terms. In Section 4 we will revisit this result by considering different levels of active, strategic, and diversification value-add.

3.4 False-Positive Probabilities

We now turn to the opposite case in which realized value-add is positive. Suppose the fund has generated 1% realized value-add relative to its benchmark, which fund management views as evidence of superior skill. An external observer may instead ask: what is the likelihood that a fund with no superior skill could nonetheless generate this level of outperformance over the evaluation period?

To address this question, we assess the risk of attributing luck to skill by calculating the corresponding false-positive probabilities across the different portfolios. For each pair of portfolios, we consider a situation in which the true value-add is zero and the observed performance spread is 1%.¹³ Table 4 reports the results using the same format as Table 3. We also report T_0^* , the number of years required to reduce the false-positive probability

¹³If the observed value-add is zero, the false-positive probability is 50% since, under $\alpha = 0$, the sampling distribution of the realized return spread is symmetric around zero. In this case, the false-positive probability provides no information for distinguishing skill from luck.

below 5%.¹⁴ Because the observed performance spread is fixed at 1% across portfolio pairs, differences in false-positive rates are driven solely by differences in tracking error.

Table 4: **False-Positive Probabilities.** This table reports the levels of value-add (true and observed), tracking error, and false-positive probabilities for a series of performance tests. Each test is based on the return spread between a portfolio and a selected benchmark. Value-add is measured as the average return spread, while tracking error is measured as the standard deviation of the return spread. The variable $P_0(T)$ denotes the probability that the realized value-add is above the observed value, conditional on the portfolio having zero true value-add, over either a one-year ($T = 1$) or 10-year ($T = 10$) investment horizon. The variable T_0^* denotes the number of years required for $P_0(T)$ to be equal to 5%. We assume that the underlying return distribution corresponds either to the 2021 capital market assumptions from T. Rowe Price or to the historical return distribution from 1998–2021 from CEM Benchmarking.

Test	Portfolio	Benchmark	Value-Add		Tracking	$P_0(T)$		T_0^*
			True	Obs.	Error	1-year	10-years	
<i>True Return Distribution = Alternative CMA</i>								
DvR	Diversified	Reference	0.00	1.00	3.92	0.40	0.21	41.78
SvD	Strategic	Diversified	0.00	1.00	2.11	0.32	0.07	12.14
AvS	Actual	Strategic	0.00	1.00	1.90	0.30	0.05	9.83
AvR	Actual	Reference	0.00	1.00	3.78	0.40	0.20	38.98
<i>True Return Distribution = Historical Distribution</i>								
DvR	Diversified	Reference	0.00	1.00	4.98	0.40	0.26	67.54
SvD	Strategic	Diversified	0.00	1.00	2.16	0.32	0.07	12.73
AvS	Actual	Strategic	0.00	1.00	1.90	0.30	0.05	9.83
AvR	Actual	Reference	0.00	1.00	3.95	0.40	0.21	42.41

For the total portfolio test *AvR*, the false-positive probability is 40% over a one-year horizon and 20-21% over a 10-year horizon. The estimates are again similar for both proxies of the true return distribution. It takes approximately 40 years for the false-positive probability to fall below 5%. These results indicate that a fund with no superior investment skill has a high probability of generating at least 1% annualized value-add over a 10-year horizon purely by chance.

Turning to adjacent portfolio pairs, the return spread between the reference and diversified portfolios exhibits the largest tracking error of the three tests. Consequently, a performance test of the diversified portfolio against the reference portfolio (*DvR*) has the highest false-positive probability, ranging from 21% to 26% over a 10-year horizon.

¹⁴ T_0^* is equal to $T_0^* = \left(\frac{\bar{z}_0^* \sigma_\epsilon}{\bar{\alpha}_{obs}} \right)^2$, where $\bar{z}_0^* = 1.645$.

By comparison, the false-positive probability is 7% when testing the strategic portfolio against the diversified portfolio (SvD) and 5% when testing the actual portfolio against the strategic portfolio (AvS).

Altogether, our calibration results highlight the substantial noise involved in distinguishing investment skill from luck for institutional funds over realistic evaluation horizons. Point estimates of performance provide only a partial view of the evidence for skill, and the false-negative and false-positive probabilities provide useful tools for quantifying and interpreting this uncertainty. These probabilities remain high over practical investment horizons, both for the total portfolio and across the individual components of the value chain, with diversification being particularly affected.

4 Optimal Benchmark Design

The limited ability of the reference portfolio test to distinguish skill from luck raises important questions for benchmark design. When is the reference portfolio the most—and least—effective benchmark for detecting skill? Can alternative benchmarks reduce false-negative risk for funds that possess genuine skill? To shed light on these questions, we develop a theoretical model based on the mean-variance framework that reduces the portfolio problem to its key parameters and essential features. The model allows us to characterize the sources and properties of noise along the value chain and conduct a series of counterfactual tests for different levels of diversification, strategic, and active value-add. These results, in turn, allow us to assess which benchmark designs are most informative for distinguishing skill from luck, depending on the properties of each source of value-add.

4.1 Modeling the Mean-Variance Framework

4.1.1 Notation and Simplifying Assumptions

The portfolio construction framework is the same as in Sections 2 and 3. For each portfolio $i \in \{R, D, S, A\}$, we solve the mean-variance optimization problem for a given set of asset classes and return forecasts, subject to a volatility constraint $\bar{\sigma}$. We then evaluate the risk-return properties of each portfolio under the true return distribution and apply the performance tests. The difference is that, rather than specifying a distinct set of CMA forecasts for each portfolio as in the calibration exercise, we directly model distortions in return forecasts; the methodology is described in the next section. We also abstract from the institutional details and do not impose portfolio constraints. These simplifications allow us to isolate the key economic forces that determine how informative different benchmarks are for evaluating fund performance.

We first set up the notation and make several simplifying assumptions. Cash is risk-free, and we normalize the risk-free rate to zero. There are no constraints on portfolio weights. The N benchmark funds have a (true) expected-return vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We assume that benchmark funds in listed equities and bonds have expected-return vector $\boldsymbol{\mu}_2$ and covariance matrix $\boldsymbol{\Sigma}_2$ and are uncorrelated with the remaining $N - 2$ benchmark funds. We denote by λ and λ_2 the maximum Sharpe ratios attainable by investing in all N benchmark funds and in the benchmark funds of listed equities and bonds, respectively. We model the fund's active management as a zero-weight portfolio with value-add α_A , volatility $\sigma_{\epsilon,A}$, and zero correlation with the benchmark funds.

Under these assumptions, the model reduces the analysis to three key parameters that shape the value-add of each source of skill: (1) the degree of diversification, captured by the ratio λ_2/λ ; (2) the extent of return-forecast misspecification, which we introduce in the next section; and (3) the level of active value-add, α_A . We vary these parameters individually, holding the remaining parameters fixed to isolate their effects on benchmark informativeness.

4.1.2 Modelling Return Forecast Distortions

A distinctive feature of our model is how we model distortions in investors' return forecasts. We assume that investors correctly perceive the covariance matrix of asset returns but may have incorrect estimates of expected returns.¹⁵ Let $\mathbf{u} = \Sigma^{-1/2}\boldsymbol{\mu}/\|\Sigma^{-1/2}\boldsymbol{\mu}\|$ denote the normalized, covariance-adjusted expected-return vector, so that $\|\mathbf{u}\| = 1$. We similarly define $\mathbf{u}_2$ for the first two asset classes.¹⁶ We assume that each portfolio $i \in \{R, D, S, A\}$ is constructed using the forecast

$$\tilde{\boldsymbol{\mu}}_i = \Sigma^{1/2}\lambda_i (\cos(a_i)\mathbf{u}_i + \sin(a_i)\mathbf{v}_i), \quad (13)$$

where $\lambda_i \in \lambda$, λ_2 , $\mathbf{u}_i \in \mathbf{u}, \mathbf{u}_2$, and $\mathbf{v}_i$ is a unit vector satisfying $\mathbf{u}_i'\mathbf{v}_i = 0$. The vector $\mathbf{v}_i$ determines the direction of the forecast distortion, while the scalar $a_i \in [0, \pi/2]$ determines its magnitude. When $a_i = 0$, investors correctly perceive expected returns, while larger values correspond to progressively greater distortions. We provide the full details of the forecast-distortion specification in the Appendix.¹⁷

The forecasts underlying the reference, diversified, strategic, and actual portfolios differ in the direction of forecast distortion, $\mathbf{v}_i$, and its magnitude, a_i . The reference portfolio uses $(\mathbf{v}_R, \bar{a})$, the diversified portfolio uses $(\mathbf{v}_D, \bar{a})$, and the strategic portfolio uses $(\mathbf{v}_S, a)$. We set $0 \leq a \leq \bar{a}$, so that the fund has greater return-forecasting ability than individual investors, reflecting its strategic skill. The actual portfolio has the same weights as the strategic portfolio and therefore shares its parameters $\mathbf{v}_S$ and a . We allow the distortion vectors to be correlated, with correlations $\rho_{R,D}^v$, $\rho_{D,S}^v$, and $\rho_{R,S}^v$, to capture common errors in investors' forecasts.

Equation (13) provides a convenient and tractable representation of the perceived expected-return vector with three useful properties. First, the specification is without

¹⁵This assumption is sensible because covariances can generally be estimated more precisely than expected returns. Relaxing this assumption is straightforward.

¹⁶Specifically, $\mathbf{u}_2$ is obtained by normalizing $\Sigma_2^{-1/2}\boldsymbol{\mu}_2$ and extending the resulting vector by appending $N - 2$ zeros.

¹⁷This specification draws on Rodriguez Dominguez (2026), who introduces a family of surrogate return signals similar to those considered here and derives their implications for portfolio efficiency.

loss of generality: any perceived expected-return vector can be uniquely decomposed into a component aligned with the true expected-return vector and an orthogonal component, so $\mathbf{u}_i' \mathbf{v}_i = 0$ is a geometric property of the decomposition rather than a restriction on estimation errors. Second, the specification preserves the overall attractiveness of the investment opportunity set: investors do not systematically over- or underestimate its attractiveness, but instead misjudge the relative attractiveness of individual opportunities. Third, forecast distortions are larger for riskier assets, consistent with the idea that expected returns are harder to estimate for assets whose prices contain more noise. We establish these properties formally in the Appendix.

4.1.3 Sources of Value-Add and Key Parameters

We obtain closed-form expressions for the value-add associated with each source of skill, each depending on a small number of key parameters. The active value-add, α_A , captures the fund's risk-adjusted returns from active management and is positive when these returns are positive. The strategic value-add,

$$\alpha_S = \bar{\sigma} \lambda (\cos(a) - \cos(\bar{a})), \quad (14)$$

is positive when the fund makes superior investment return forecasts ($a < \bar{a}$), since the cosine function takes values between 0 and 1 and is strictly decreasing over the interval $[0, \pi/2]$. Finally, the diversification value-add,

$$\alpha_D = \bar{\sigma} \lambda \cos(\bar{a}) \left(1 - \frac{\lambda_2}{\lambda}\right), \quad (15)$$

is positive when the fund generates diversification benefits ($\lambda_2/\lambda < 1$).

In the sections that follow, we vary each source of value-add from 0 to 1.5% per year by changing: (i) the active value-add, α_A ; (ii) the fund's degree of forecast error, a , which affects α_S ; and (iii) the Sharpe ratio of the portfolio of N benchmark funds, λ , which affects α_D .

4.1.4 Model Parametrization

We hold the remaining parameters fixed throughout the analysis. Similar to Section 3, we set the volatility target $\bar{\sigma}$ to 10% and the tracking error of active management, $\sigma_{\epsilon,A}$, to 1.9%. We set the Sharpe ratio of the optimal passive portfolio of listed stocks and bonds, λ_2 , to 0.4, in line with typical estimates (Doeswijk et al., 2020). We set the degree of forecast error among individual investors, $\bar{a}$, to $0.6\pi/2$. This ensures that varying the fund's degree of forecast error, a , from 0 to $\bar{a}$ generates values of α_S in the 0-1.5% range.

We consider two sets of values for the correlations $\rho_{R,D}^v$, $\rho_{R,S}^v$ and $\rho_{D,S}^v$. The first case is designed to match the estimates from the calibrated example in Section 3. Specifically, we set $\rho_{R,D}^v = 0.9$, $\rho_{R,S}^v = 0.9$, and $\rho_{D,S}^v = 0.9$, reflecting a high degree of commonality in investors' forecasts.¹⁸ We refer to this case as the high-correlation case.

As a contrasting case, we reduce the correlations $\rho_{R,D}^v$ and $\rho_{R,S}^v$ to 0.1, while keeping $\rho_{D,S}^v$ at 0.9. This captures a situation in which individual investors' forecasts differ substantially from those of institutional funds. This may arise, for example, when the reference portfolio is set by the board in ways that are largely unrelated to the diversified and strategic portfolios. We refer to this case as the low-correlation case. Examining both cases allows us to assess the effectiveness of different benchmarks across a range of plausible scenarios.

4.2 Sources and Properties of Noise Along the Value Chain

4.2.1 Portfolio Optimization as a Source of Tracking Error

The model allows us to characterize how different layers of portfolio optimization generate tracking error along the value chain. As explained in Section 2, the tracking error between two portfolios depends on their volatilities and their correlation. In the model, all portfolios have the same volatility $\bar{\sigma}$ by design, except for the fund's actual portfolio, whose volatility is higher due to the tracking error $\sigma_{\epsilon,A}$. Thus, with portfolio volatilities

¹⁸In this calibrated case, we find that when each value-add is 0.5% and the total value-add is therefore 1.5%, the probability of obtaining a false-negative is approximately 15%, similar to what we observe in Table 3.

otherwise held fixed, imperfect correlation is the primary source of tracking error.¹⁹

To study correlations, it is helpful to begin with the special case in which investors correctly forecast expected returns (i.e. $a = \bar{a} = 0$). This allows us to isolate the tracking error that arises from portfolio optimization in the diversification step. We then introduce forecast uncertainty and show how it generates an additional component of tracking error in both the diversification and strategic steps.

Case Without Forecast Uncertainty In the case where investors correctly forecast expected returns, the correlation between each pair of adjacent portfolios is given by:

$$\rho_{i,j} = \begin{cases} \frac{\lambda_2}{\lambda} & \text{for } (i,j) = (R,D) \\ 1 & \text{for } (i,j) = (D,S) \\ \frac{\bar{\sigma}}{\sqrt{\bar{\sigma}^2 + \sigma_{\epsilon,A}^2}} & \text{for } (i,j) = (S,A) \end{cases} \quad (16)$$

The correlation between the reference and diversified portfolios, $\rho_{R,D}$, depends on the magnitude of the diversification benefits associated with N assets. As diversification benefits increase (i.e., as the ratio λ_2/λ decreases), the reference and diversified portfolios become more distinct and, consequently, less correlated. In contrast, the diversified and strategic portfolios have perfect correlation because, in the special case where $a = \bar{a} = 0$, both portfolios are identical. Finally, the correlation between the strategic and actual portfolios, $\rho_{S,A}$, depends on the amount of tracking error associated with active value-add, $\sigma_{\epsilon,A}$.

The key difference between the mechanisms driving $\rho_{R,D}$ and $\rho_{S,A}$ in Equation (16) is portfolio re-optimization. In both the diversification and active steps, the fund has access to a new investment opportunity that is uncorrelated with the existing assets,

¹⁹Our approach is conservative in that we shut down one potential source of tracking error by assuming that investors share the same volatility forecasts. If investors instead have different forecasts of volatility, they may construct portfolios with different realized volatilities, generating additional tracking error. For example, if individual investors perceive volatility to be lower than institutional investors do, they may ultimately take on more risk. Thus, tracking error can arise not only from imperfect correlation between portfolios but also from differences in their realized volatilities.

suggesting that the two correlations might behave similarly. However, the strategic and actual portfolios assign the same weights to the N benchmark funds, whereas the reference and diversified portfolios assign different weights to the first two benchmark funds. Specifically, when constructing the diversified portfolio, the addition of the remaining $N - 2$ asset classes leads to a re-optimization of the weights on the first two. This re-optimization moves the diversified portfolio further away from the reference portfolio and reduces their correlation. In the Appendix, we show that if the actual portfolio were similarly re-optimized after incorporating the return from active management, its correlation with the strategic portfolio would likewise decline.

This portfolio optimization channel has two implications for the informativeness of the performance test. First, portfolio optimization generates tracking error during the diversification step. Second, the tracking error increases with the level of value-add. As Equations (15) and (16) show, greater value-add from the diversification channel leads to greater portfolio optimization and, in turn, higher tracking error.

General Case With Forecast Uncertainty We next examine the correlation structure in the more general case where investors have incorrect forecasts about expected returns:

$$\rho_{i,j} = \begin{cases} \cos^2(\bar{a}) \frac{\lambda_2}{\lambda} + \sin^2(\bar{a}) \rho_{R,D}^v & \text{for } (i, j) = (R, D) \\ \cos(\bar{a}) \cos(a) + \sin(\bar{a}) \sin(a) \rho_{D,S}^v & \text{for } (i, j) = (D, S) \\ \frac{\bar{\sigma}}{\sqrt{\bar{\sigma}^2 + \sigma_{\epsilon,A}^2}} & \text{for } (i, j) = (S, A) \end{cases} \quad (17)$$

Misspecified forecasts do not affect the correlation between the strategic and actual portfolios, $\rho_{S,A}$, because both portfolios are constructed using the same forecasts; the only difference between them is the addition of the uncorrelated active return. In contrast, misspecified forecasts introduce an additional source of separation between the reference and diversified portfolios, $\rho_{R,D}$, and between the diversified and strategic portfolios, $\rho_{D,S}$. As the degree of forecast distortion increases, greater weight is placed on the correlations $\rho_{R,D}$ and $\rho_{D,S}$, making the commonality of forecast errors across investors increasingly

important for the resulting portfolio correlations. When those forecast errors are less correlated across investors, their resulting portfolios are more distant from one another, leading to lower correlations between their returns.

These results highlight another channel through which portfolio optimization under forecast uncertainty affects tracking error: a fund with greater ability to forecast expected returns deviates further from its benchmark, generating higher value-add but lower correlation and, consequently, higher tracking error.

Cumulative Effects of Portfolio Optimization From Equation (17), we obtain the correlation between the actual and reference portfolios:

$$\rho_{A,R} = \left(\cos(\bar{a}) \cos(a) \frac{\lambda_2}{\lambda} + \sin(\bar{a}) \sin(a) \rho_{R,S}^v \right) \frac{\bar{\sigma}}{\sqrt{\bar{\sigma}^2 + \sigma_{\epsilon,A}^2}}. \quad (18)$$

The correlation between the reference and actual portfolios reflects all sources of imperfect correlation along the value chain. Each successive stage of portfolio optimization lowers the correlation with the reference portfolio, thereby increasing tracking error at the total portfolio level.

4.2.2 How Optimization-Induced Tracking Error Affects Test Informativeness

To illustrate how optimization-induced tracking error affects the informativeness of the corresponding performance test, we consider a series of counterfactual scenarios in which the fund's value-add comes from one source at a time, with value-add from all other sources set to zero. For each scenario, we run the performance test for the relevant pair of adjacent portfolios and estimate the false-negative probability over a 10-year evaluation period, using zero as the threshold for observed value-add $\bar{\alpha}_{obs}$.

Figure 2 plots the 10-year false-negative probability, $P(10)$, as a function of the active value-add (α_A), the strategic value-add (α_S), and the diversification value-add (α_D), which we vary from 0% to 1.5%. We consider both the high-correlation and low-

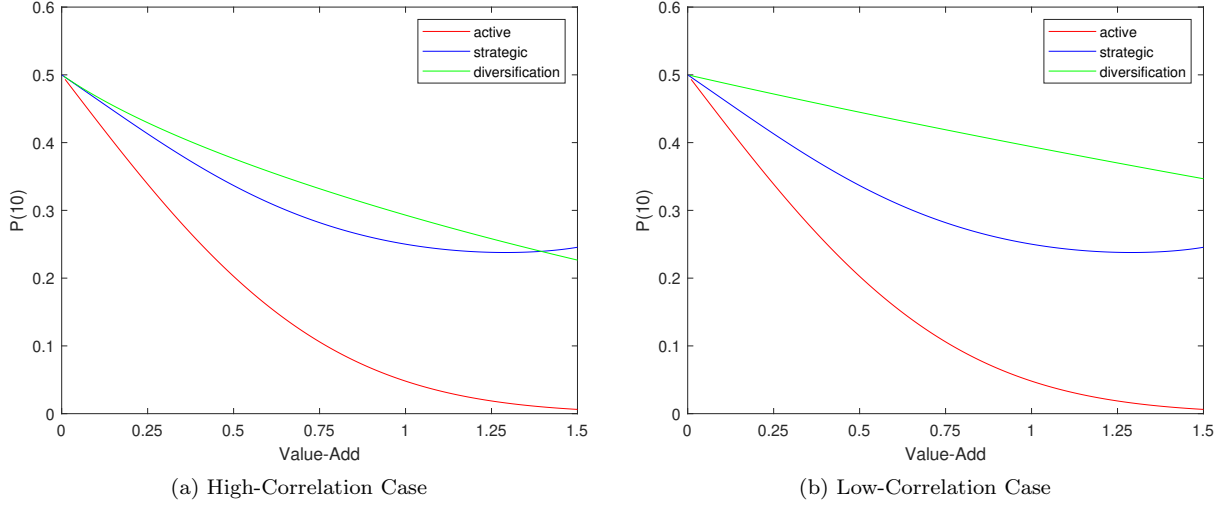


Figure 2: False-Negative Probability for Each Source of Value-Add. This figure plots the 10-year false-negative probability, $P(10)$, for each value-add component as a function of the level of value-add, holding the other value-adds equal to zero. Each panel corresponds to a different specification of the correlations between forecast errors. The high-correlation case (Panel A) sets $\rho_{R,D}^v = 0.9$, $\rho_{R,S}^v = 0.9$, and $\rho_{D,S}^v = 0.9$, while the low-correlation case (Panel B) sets $\rho_{R,D}^v = 0.1$, $\rho_{R,S}^v = 0.1$, and $\rho_{D,S}^v = 0.9$. All value-add measures and $P(10)$ are reported in percentage terms.

correlation cases introduced in Section 4.1.4. The downward-sloping lines indicate that $P(10)$ is generally decreasing with the level of value-add, as expected. As the true value-add approaches zero, the false-negative probability approaches 50%.

Panel A of Figure 2 depicts the high-correlation case in which forecast errors are highly correlated across portfolios. We draw two insights from the figure. First, $P(10)$ is higher for diversification value-add than for active value-add and declines less rapidly as diversification value-add increases. For example, increasing diversification value-add from 0% to 1.5% per year reduces $P(10)$ from 50% to 23%, whereas the same increase in active value-add reduces $P(10)$ from 50% to 0.6%. This difference in sensitivity reflects the portfolio optimization channel discussed earlier: higher diversification value-add results in greater deviation from the reference portfolio, which in turn generates higher tracking error. Consequently, the false-negative probability for the DvR test remains high even at substantial levels of diversification value-add.

Second, $P(10)$ is not monotonically decreasing in strategic value-add, α_S ; instead, it begins to increase once strategic value-add exceeds 1.29%. This reversal reflects the

interaction between portfolio optimization and forecast uncertainty. As funds with larger strategic value-add increasingly tilt away from the diversified portfolio, they generate higher tracking error, which eventually offsets the benefits of additional value-add. This effect becomes more pronounced as the magnitude of the portfolio tilt increases.

In Panel B of Figure 2, we consider the low-correlation case in which individuals' forecasts are weakly correlated with those of institutional funds. The tracking error of the diversified portfolio relative to the reference portfolio increases materially, resulting in a higher false-negative probability. For example, even when the diversification value-add is as high as 1.5%, $P(10)$ remains around 35%. This result illustrates how lower correlation among forecast errors increases false-negative rates for performance tests at upstream stages of the value chain.

4.3 Which Benchmark Portfolio Minimizes False-Negative Risk?

We next examine when alternative benchmarks are more effective than the reference portfolio at detecting investment skill. Starting from the fund's actual portfolio, we compare the reference, diversified, and strategic portfolios as potential benchmarks under various scenarios.

4.3.1 A Fund With Active Skill Only

We first examine a setting in which the fund's only source of value-add is active management. In Figure 3 we vary α_A from 0 to 1.5% while setting α_D and α_S to zero, and report the value of $P(10)$ for three tests: the incremental test that compares the actual and strategic portfolios (AvS); the combined test that compares the actual and diversified portfolios (AvD); and the total portfolio test that compares the actual and reference portfolios (AvR). We report the results for both the high- and low-correlation cases.

Starting with the high-correlation case in Panel A, the high values of $P(10)$ for the AvD and AvR tests indicate that the false-negative probability increases when we run combined performance tests. This result is intuitive because, in this particular exercise,

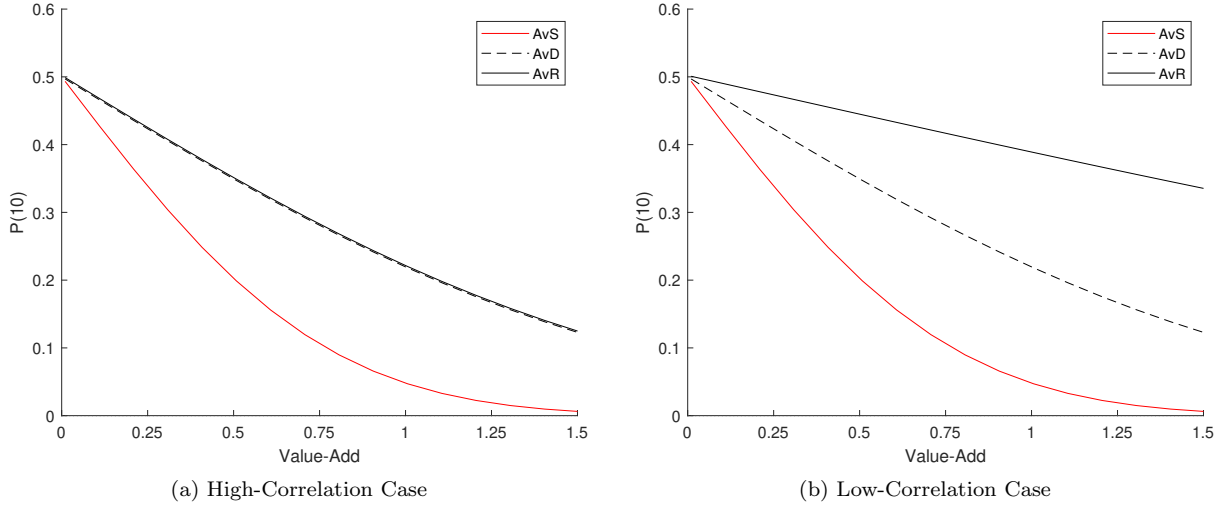


Figure 3: Performance Tests with Alternative Benchmarks for a Fund with Active Skill Only. This figure plots the 10-year false-negative probability, $P(10)$, for several performance tests for a fund with varying levels of active value-add and no diversification and strategic value-adds. The tests include: the individual test comparing the active and strategy portfolios (AvS), the combined test comparing the active and diversified portfolios (AvD), and the total portfolio test comparing the actual and reference portfolios (AvR). Each panel corresponds to a different specification of the correlations between forecast errors. The high-correlation case (Panel A) sets $\rho_{R,D}^v = 0.9$, $\rho_{R,S}^v = 0.9$, and $\rho_{D,S}^v = 0.9$, while the low-correlation case (Panel B) sets $\rho_{R,D}^v = 0.1$, $\rho_{R,S}^v = 0.1$, and $\rho_{D,S}^v = 0.9$. All value-add measures and $P(10)$ are reported in percentage terms.

the combined tests introduce noise without incorporating any additional value-add.

In Panel B, we consider the low-correlation case. The weaker correlation between the reference and diversified portfolios' forecast errors implies that the total portfolio test has an even weaker ability to detect value-add from active management. For example, when the active value-add is 1.5% per year, the AvR combined test produces a false-negative probability that is 32 percentage points higher than that of the AvS test.

4.3.2 A Fund with Diversification and Strategic Skills

We next consider a fund with superior diversification and strategic skills and examine how the magnitude of these skills affects the choice of benchmark. In Figure 4, we set α_A to 0.29%, as in the calibrated example of Section 3, and vary the other value-adds in equal proportion. That is, we consider different values of the sum $\alpha_D + \alpha_S$, with α_D and α_S increasing equally. As before, we report the value of $P(10)$ for three tests - AvS , AvD , and AvR - under both the high- and low-correlation cases.

In the high-correlation case (Panel A), when strategic and diversification value-adds exceed 0.40% in total, the combined tests have a smaller false-negative probability than the AvS test. This indicates that combined performance tests are better able to detect investment skill when the fund generates sufficient value from multiple sources. We also find that the combined test AvD has a smaller false-negative probability than the total portfolio test AvR . When diversification value-add generates particularly high tracking-error noise, excluding this component improves the informativeness of the performance test. The AvD test therefore removes the noise associated with diversification while retaining the signals from strategic and active value-add.

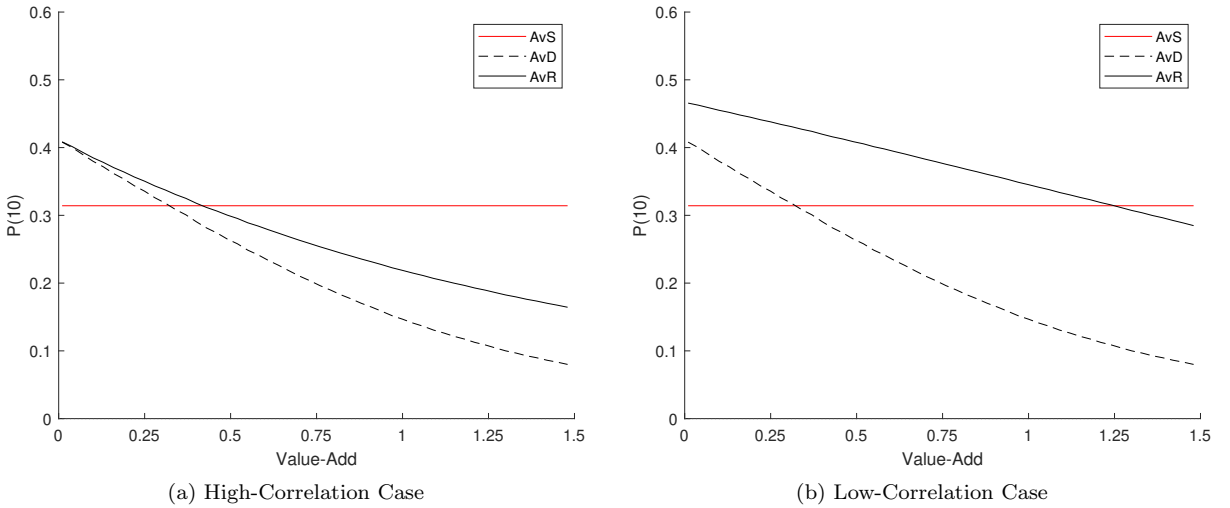


Figure 4: Performance Tests with Alternative Benchmarks for a Fund with Multiple Skills. This figure plots the 10-year false-negative probability, $P(10)$, for several performance tests for a fund with varying levels of diversification and strategic value-adds and active value-add set to 0.29%. The tests include: the individual test comparing the active and strategy portfolios (AvS), the combined test comparing the active and diversified portfolios (AvD), and the total portfolio test comparing the actual and reference portfolios (AvR). Each panel corresponds to a different specification of the correlations between forecast errors. The high-correlation case (Panel A) sets $\rho_{R,D}^v = 0.9$, $\rho_{R,S}^v = 0.9$, and $\rho_{D,S}^v = 0.9$, while the low-correlation case (Panel B) sets $\rho_{R,D}^v = 0.1$, $\rho_{R,S}^v = 0.1$, and $\rho_{D,S}^v = 0.9$. All value-add measures and $P(10)$ are reported in percentage terms.

In the low-correlation case (Panel B), the false-negative probability of the total portfolio test AvR increases because diversification value-add generates greater tracking error relative to the reference portfolio. Consequently, AvR dominates AvS only when strategic and diversification value-adds exceed 1.25% in total. In contrast, the false-negative probability of AvD is unchanged, as this test is unaffected by the lower correlation between

the forecast errors of individual and institutional investors' portfolios.²⁰

These results altogether show how the optimal benchmark choice for minimizing false-negative risk depends on the signal-to-noise ratio associated with each source of value-add. A fund confident in its ability to generate value-add from all three sources and facing favorable signal-to-noise ratios across these sources may choose the reference portfolio as its benchmark. A fund with strong strategic and active value-add but highly noisy diversification value-add may instead choose the diversified portfolio as its benchmark. By contrast, a fund with limited ability to generate diversification and strategic value-add relative to the tracking error they introduce may choose the strategic portfolio as its benchmark.

5 Implications for Practice and Policy

Our findings have important implications for fund managers, board members, policy-makers, and journalists when interpreting performance results and selecting appropriate benchmarks.

Noisy Performance Test Separating skill from luck at the total portfolio level is extremely difficult, and results from performance tests of a fund's total portfolio should be interpreted with caution. The high rates of false negatives and false positives imply that these tests frequently misclassify the skill of a fund's management team, even over extended horizons. In addition, noise in total portfolio performance makes it difficult to identify the source of investment skill.

Importantly, our results do not call into question the value of benchmarking; rather, they highlight the limitations of existing benchmarks for inferring managerial skill and the need for improved methods that better distinguish skill from luck. Point estimates of performance alone provide only a partial view of the evidence for skill. Reporting false-negative and false-positive probabilities alongside the point estimates can help quantify

²⁰If $\rho_{D,S}^v$ decreased, the tracking error between the strategic and actual portfolios would then increase, making the AvD test less effective.

and interpret the uncertainty surrounding these inferences.

Reference portfolios remain valuable tools for fund management and governance. They provide fund management teams with a framework for aligning objectives and help fund boards calibrate the appropriate level of risk by linking risk to a concrete asset allocation rather than considering volatility in isolation. They also provide a useful counterfactual for evaluating how investors might optimally allocate capital in the absence of the fund. However, their usefulness for setting objectives and risk does not imply that they provide a reliable basis for determining whether realized value-add reflects managerial skill or luck.

Combining Skill Sets Into a Single Performance Test Combining multiple sources of value-add into a single performance test has both benefits and limitations. A combined test has greater ability to detect investment skill when each source of value-add generates a strong signal relative to its associated tracking error. However, if one component introduces substantial tracking error relative to its value-add, the resulting noise can make the combined test less informative.

This result is important because the tracking error associated with diversification value-add is particularly large.²¹ This is noteworthy given that diversification value-add is among the most well-established results in academic finance, dating back to Markowitz’s seminal work on portfolio selection (Markowitz, 1952).

“The hypothesis (or maxim) that the investor does (or should) maximize discount return must be rejected. If we ignore market imperfections, the foregoing rule never implies that there is a diversified portfolio which is preferable to all non-diversified portfolios. Diversification is both observed and sensible; a rule of behavior which does not imply the superiority of diversification must be rejected both as a hypothesis and as a maxim.”

— Harry Markowitz

²¹This issue may be especially salient following the strong performance of listed equities in recent years, which can make the value-add from diversification into other assets more difficult to distinguish from noise.

The high level of tracking error associated with diversification therefore raises an important question regarding benchmark selection: if diversification value-add is well established but its associated tracking error is large, should funds exclude it from the performance test by using the diversified portfolio as the benchmark rather than the reference portfolio, as our theoretical results suggest? More generally, should sources of value-add that are well established but difficult to measure be incorporated into the benchmark rather than evaluated through the performance test?

Total Portfolio Approach There is an inherent tension between the total portfolio approach and the use of the reference portfolio as a performance benchmark. The total portfolio approach is designed to allow a fund’s management team to exploit its investment expertise and optimize the portfolio. However, portfolio optimization creates tracking error, and for a benchmark portfolio to effectively measure skill, it must remain closely aligned with the actual portfolio in terms of low volatility differential and high correlation. Consequently, as a fund’s actual portfolio deviates further from the reference portfolio, the reference portfolio becomes a noisier measure of investment skill.

Ex Ante Performance Evaluation Our finding that performance tests are highly noisy at all stages of the value chain suggests that performance evaluation should consider complementary approaches that examine the quality of investment processes ex ante. For example, external auditors could assess whether funds have efficient cost structures, adhere to their governance processes, and implement their investment strategies consistently. Such audits may provide additional insights into the quality of a fund’s investment process.

6 Conclusion

In this paper, we develop a parsimonious framework to establish a key point: when the reference portfolio is used to evaluate an institutional fund’s total portfolio, performance tests are highly noisy and can generate high rates of false positives and false negatives

over practical evaluation horizons. This noise arises for several reasons. First, portfolio optimization creates tracking error as funds deviate from their reference portfolios in pursuit of value-add. Second, uncertainty about expected returns can lead to further differences between optimized and reference portfolios. As a result, some sources of value-add, particularly diversification, generate substantial tracking-error noise relative to the signal they provide, making it difficult to reliably evaluate a fund’s total portfolio performance.

Our analysis provides tools to quantify and interpret the uncertainty in reference portfolio test results and offers guidance on improving benchmark selection. Board members and trustees of institutional funds should recognize a tension between portfolio optimization and benchmarking: the more a fund optimizes away from its benchmark to create value, the less informative that benchmark becomes for performance evaluation. When choosing the appropriate benchmark, board members and trustees should reflect on the multiple sources of value-add a fund seeks to exploit, the value they are expected to generate, and the tracking error they create. The optimal benchmark may differ across funds, and the associated trade-offs should be clearly articulated.

Future research could extend our analysis in several directions. One avenue is to investigate how additional features of fund returns, such as lagged and smoothed private-asset returns, affect the reliability of performance evaluation (Beath and Flynn, 2020; Baz et al., 2022; Boyer et al., 2023). Another is to develop performance tests tailored to specific sources of value-add. For example, the value of diversification benefits could be better assessed by using longer historical time-series outside the observed sample period (Lynch and Wachter, 2013). The value of active performance in unlisted markets can be sharpened by adjusting for smoothing and lagging (Beath et al., 2022). A third is to develop approaches for evaluating investment processes *ex ante* rather than relying solely on realized outcomes *ex post*. Such approaches are currently rarely used in practice, suggesting an unmet demand for tools that can evaluate investment processes before their outcomes are realized.

References

- Andonov, A., S. J. Coutts, A. S. Gonçalves, J. Loudis, and A. Rossi (2026). Subjective beliefs and the portfolio allocations of institutional investors. Working Paper.
- Back, K. E. (2017). *Asset Pricing and Portfolio Choice Theory* (2nd ed.). Oxford: Oxford University Press.
- Basak, S. (2000). A model of dynamic equilibrium asset pricing with heterogeneous beliefs and extraneous risk. *Journal of Economic Dynamics and Control* 24(1), 63–95.
- Baz, J., J. M. Davis, L. Han, and C. Stracke (2022). The value of smoothing. *Journal of Portfolio Management* 48(9), 73–85.
- Beath, A. and C. Flynn (2020). Benchmarking the performance of private equity portfolios of the world’s largest institutional investors: A view from CEM Benchmarking. *The Journal of Investing* 30(1), 67–87.
- Beath, A. D., S. Betermier, M. Van Bragt, Y. Liu, and Q. Spehner (2022). Green urban development: The impact investment strategy of Canadian pension funds. *Journal of Sustainable Real Estate* 14(1), 75–94.
- Beath, A. D. and C. Flynn (2023). Asset allocation and fund performance of defined benefit pension funds in the united states, 1998–2021. Technical report, CEM Benchmarking Inc., Toronto, ON, Canada.
- Beath, A. D., C. Flynn, R. Jethalal, and M. Reid (2022, February). A case for scale: How the world’s largest institutional investors leverage scale to deliver real outperformance. Technical report, CEM Benchmarking.
- Blake, D. and A. Timmermann (2002). Performance benchmarks for institutional investors: Measuring, monitoring and modifying investment behaviour. In J. Knight and S. Satchell (Eds.), *Performance Measurement in Finance: Firms, Funds and Managers*, Chapter 5, pp. 108–141. Oxford: Butterworth-Heinemann.
- Boyer, B. H., T. D. Nadauld, K. P. Vorkink, and M. S. Weisbach (2023). Discount-rate risk in private equity: Evidence from secondary market transactions. *Journal of Finance* 78(2), 835–885.

- Brinson, G. P., L. R. Hood, and G. L. Beebower (1986). Determinants of portfolio performance. *Financial Analysts Journal* 42(4), 39–44.
- Brinson, G. P., B. D. Singer, and G. L. Beebower (1991). Determinants of portfolio performance II: An update. *Financial Analysts Journal* 47(3), 40–48.
- Doeswijk, R. Q., T. Lam, and L. Swinkels (2020). Historical returns of the market portfolio. *The Review of Asset Pricing Studies* 10(3), 521–567.
- Elkamhi, R. and J. S. H. Lee (2025). Total portfolio approach. *The Journal of Portfolio Management* 51(8), 109–127.
- Grinblatt, M. and S. Titman (1995). Performance evaluation. In R. A. Jarrow, V. Maksimovic, and W. T. Ziemba (Eds.), *Handbooks in Operations Research and Management Science: Finance*, Volume 9, pp. 581–609. Amsterdam: Elsevier.
- Grinold, R. C. and R. N. Kahn (2000). *Active Portfolio Management: A Quantitative Approach for Providing Superior Returns and Controlling Risk* (2nd ed.). New York: McGraw-Hill.
- Ibbotson, R. G. and P. D. Kaplan (2000). Does asset allocation policy explain 40, 90, or 100 percent of performance? *Financial Analysts Journal* 56(1), 26–33.
- Kloepfer, J., A. Lozinski, and K. Machiz (2021). 2021 capital markets assumptions webinar.
- Lustig, Y., T. Poullaouec, L. Yura, and P. Austin (2021). Capital market assumptions five-year perspective 2021. Technical report, T. Rowe Price. 2021 Capital Market Assumptions.
- Lynch, A. W. and J. A. Wachter (2013). Using samples of unequal length in generalized method of moments estimation. *Journal of Financial and Quantitative Analysis* 48, 277–307.
- Markowitz, H. (1952). Portfolio selection. *Journal of Finance* 7(1), 77–91.
- Merton, R. C. (1980). On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics* 8(4), 323–361.
- Raymond, D. (2009). Integrating goals, structure, and decision-making at the canada pension plan investment board. *Rotman International Journal of Pension Manage-*

ment 2(1), 22–29.

Rodriguez Dominguez, A. (2026). Is causality necessary for efficient portfolios? a computational perspective on predictive validity and model misspecification. *SSRN Working Paper*.

Shen, S., D. Walker, and G. Rubin (2026). Measuring what matters: Evaluating the total portfolio approach. CPP Investments, Insights Institute.

Wermers, R., B. Singer, and B. R. Fischer (2026). *Performance Evaluation and Attribution Volume One: Asset Pricing and Models* (2 ed.). Academic Press.

Appendix A CMA and Historical Return Data

Table A.1 reports the CMA and historical return data. We assume that the return data represent a passive exposure to each asset class and are net of costs. The data include the annual arithmetic average return $\bar{r}_a$, geometric return $\bar{r}_g$, and volatility σ for each asset class. We also report the average correlation of each asset class to equity ($\langle \rho_e \rangle$), bond ($\langle \rho_b \rangle$), and real estate ($\langle \rho_r \rangle$). The angular braces $\langle X \rangle$ indicate the equally-weighted average of X . For some asset classes and for some providers, CMAs are not available in an obviously comparable way. This requires the use of either an estimate based on similar asset classes or a default value. Estimates are indicated in Table A.1 in *italics* and explained below.

Callan Institute CMA We obtain data from Kloepper et al. (2021). No information is provided for listed real estate, so we use private real estate as a proxy. Average returns and volatilities are taken from page 9 of the document, while correlations are taken from page 43.

T. Rowe Price CMA We obtain data from Lustig et al. (2021). Average returns are taken from page 8 of the document, while volatilities and correlations are taken from pages 10-11. No information is provided for long-duration bonds, so we assume long-

duration bonds are a 50/50 mixture of long-duration US treasuries and long-duration US credit.

Because the data report compounded geometric returns r_g rather than arithmetic returns, we recover the corresponding annualized arithmetic returns as follows. We first express the annual geometric return in terms of the average daily log return, $\mu = \log(1 + r_a)$:

$$1 + r_g = \exp \left(d\langle\mu\rangle + \frac{d}{2}\sigma_\mu^2 \right), \quad (19)$$

where $d = 250$ is the number of trading days per year and σ_μ^2 is the variance of daily log returns. We assume that daily arithmetic returns r_a are normally distributed, with mean $\langle r_a \rangle$ and standard deviation $\sigma_d = \sigma_y/\sqrt{d}$. This assumption determines the higher-order moments of r_a . We then use a fifth-order Taylor expansion of $\log(1 + r_a)$ to express $\langle\mu\rangle$ and σ_μ^2 in terms of the moments of r_a . We solve the resulting equation for $\langle r_a \rangle$ and annualize it by multiplying by d .

CEM Historical Return Data We use historical performance data for a sample of approximately 200 U.S. public- and corporate-sector defined-benefit pension funds over the period 1998–2021 from CEM Benchmarking (Beath and Flynn, 2023). Average returns correspond to the arithmetic net returns reported in Table ES6. Volatility is measured using the reported standard deviation metric in Table ES6, which captures the dispersion of average annual asset-class returns across funds. This measure does not incorporate any additional volatility arising from idiosyncratic active management. Correlations are based on the estimates reported in Table ES3. These estimates all account for reporting lags in private markets. For cash, we set the arithmetic average return to 1%, the volatility to 0.5%, and the correlations with other asset classes to zero.

Active Management Value-Add Data We obtain estimates of net value-add from the CEM Benchmarking database of institutional funds over the period 1992–2020 (Beath et al., 2022). We consider a fund with a net value-add of 0.29% and tracking error of 1.90%, corresponding to the average 1-year net value-add and tracking error among funds

with assets under management above USD 10 billion (Exhibit 1). Estimates of value-add are relative to the policy portfolio, which reflects the funds' SAA decision and corresponds to the strategic portfolio in our setting.

Table A.1: **Asset-Class Data Summary.** This table reports the data on asset-class return forecasts used in our calibration analysis. The data include: (1) 2021 capital market assumption data adapted from publicly available reports from the Callan Institute (Kloepfer et al., 2021) and T. Rowe Price (Lustig et al., 2021); and (2) historical data from US defined-benefit pension funds from CEM Benchmarking between 1998 and 2021 (Beath and Flynn, 2023). The table reports annual arithmetic average returns $\bar{r}_a$, geometric average returns $\bar{r}_g$, volatility σ , and the average correlations to equity ρ_e , bonds ρ_b , and real estate ρ_r . Data in *italics* indicates our use of calculation to obtain the estimate. – indicates data not available and our use of a default. n/a indicates data is not applicable. All data estimates are assumed to be net of costs.

Provider	Horizon	Stat.	Asset Class								
			Large Cap US Stock	Small Cap US Stock	Non-US Stock	Private Equity	Agg-regate Bonds	Long Duration Bonds	Private Real Estate	Listed Real Estate	Cash
• CMA report title											
- Estimation notes	(years)										
The Callan Institute	10	$\bar{r}_a$	0.0785	0.0875	0.0870	0.1150	0.0180	0.0230	0.0660	–	0.0100
• <i>2021 Capital Market Assumptions</i>		$\bar{r}_g$	0.0650	0.0670	0.0680	0.0800	0.0175	0.0180	0.0575	–	0.0100
- Listed real estate data not provided in source; data used in our calculation has been assumed equal to unlisted real estate.		σ	0.1770	0.2130	0.2070	0.2780	0.0375	0.1035	0.1410	–	0.0090
		$\langle \rho_e \rangle$	0.8367	0.8200	0.7967	0.7800	-0.1325	0.0600	0.6600	–	-0.0600
		$\langle \rho_b \rangle$	-0.0050	-0.0250	-0.0150	-0.1000	0.8800	0.8800	0.0300	–	0.0800
		$\langle \rho_r \rangle$	0.7100	0.6600	0.6700	0.6000	-0.0400	0.1000	1.0000	–	0.0000
T. Rowe Price	5	$\bar{r}_a$	<i>0.0602</i>	<i>0.0756</i>	<i>0.0850</i>	<i>0.0914</i>	<i>0.0115</i>	<i>-0.0042</i>	<i>0.0445</i>	<i>0.0707</i>	<i>0.0060</i>
• <i>2021 Capital Market Assumptions: Five-Year Perspective</i>		$\bar{r}_g$	0.0470	0.0520	0.0680	0.0620	0.0110	<i>-0.0105</i>	0.0380	0.0460	0.0060
		σ	0.1690	0.2230	0.1960	0.2500	0.0320	<i>0.1125</i>	0.1200	0.2270	0.0080
- $\bar{r}_a$ estimated from $\bar{r}_g$ assuming normally dist. daily returns.		$\langle \rho_e \rangle$	0.9000	0.8333	0.9000	0.8333	-0.2000	<i>-0.1875</i>	0.3750	0.7250	-0.0250
		$\langle \rho_b \rangle$	-0.1750	-0.2500	-0.1250	-0.2250	0.8000	<i>0.8000</i>	-0.2000	0.0250	0.0000
		$\langle \rho_r \rangle$	0.6000	0.5500	0.5000	0.5500	-0.1000	<i>-0.0750</i>	0.5000	0.5000	-0.0500
CEM Benchmarking	n/a	$\bar{r}_a$	0.1037	0.1083	0.0897	0.1244	0.0548	0.0798	0.0861	0.1090	0.0100
• <i>Asset allocation and fund performance of defined benefit pension funds in the United States, 1998-2021</i>		$\bar{r}_g$	0.0885	0.0915	0.0687	0.1082	0.0540	0.0765	0.0764	0.0942	0.0100
		σ_y	0.1701	0.1831	0.2006	0.1827	0.0423	0.0843	0.1419	0.1707	0.0050
		$\langle \rho_e \rangle$	0.8900	0.9000	0.8733	0.8767	0.0950	-0.2475	0.5425	0.5700	0.0000
		$\langle \rho_b \rangle$	-0.0800	-0.1050	-0.1000	0.0250	0.7500	0.7500	0.1000	0.1400	0.0000
		$\langle \rho_r \rangle$	0.5250	0.5850	0.5400	0.5750	0.2750	-0.0350	0.9000	0.9000	0.0000

Appendix B Model Derivation

In this section, we derive the formulas used for our analysis in Section 4.

Appendix B.1 Assets and Portfolios

Assets The asset universe comprises N asset classes and a risk-free asset, whose rate is normalized to zero. Each asset class $n \in \{1, \dots, N\}$ has a representative benchmark fund with a stochastic return. Let

$$\mathbf{r} \in \mathbb{R}^N, \quad \boldsymbol{\mu} \in \mathbb{R}^N, \quad \boldsymbol{\Sigma} \in \mathbb{R}^{N \times N}, \quad \boldsymbol{\Sigma} \succ 0,$$

respectively correspond to the realized return, expected return, and covariance matrix of the asset classes. Listed equities and bonds constitute the first two asset classes, with expected return vector $\boldsymbol{\mu}_2$ and covariance matrix $\boldsymbol{\Sigma}_2$. The remaining $N - 2$ asset classes have expected return vector $\boldsymbol{\mu}_{N-2}$ and covariance matrix $\boldsymbol{\Sigma}_{N-2}$. We assume listed equities and bonds are uncorrelated with the remaining asset classes. Consequently, the expected return vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ can be written as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_{N-2} \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_2 & 0 \\ 0 & \boldsymbol{\Sigma}_{N-2} \end{pmatrix}. \quad (20)$$

Maximum Sharpe ratio We show that the variable $\lambda = \sqrt{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}$ corresponds to the maximum Sharpe ratio of a portfolio of the N benchmark funds with a fixed portfolio volatility $\bar{\sigma}$.

The vector of portfolio weights $\mathbf{w}$ that maximizes the portfolio's expected return subject to the volatility $\bar{\sigma}$ is given by:

$$\max_{\mathbf{w}} \quad \boldsymbol{\mu}'\mathbf{w} \quad \text{s.t.} \quad \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} = \bar{\sigma}^2. \quad (21)$$

The Lagrangian is

$$\mathcal{L}(\mathbf{w}, \gamma) = \boldsymbol{\mu}'\mathbf{w} - \frac{\gamma}{2} (\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} - \bar{\sigma}^2). \quad (22)$$

where γ is the Lagrange multiplier. The first-order condition with respect to $\mathbf{w}$ is

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \boldsymbol{\mu} - \gamma \boldsymbol{\Sigma} \mathbf{w} = 0. \quad (23)$$

Hence,

$$\mathbf{w} = \frac{1}{\gamma} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}. \quad (24)$$

Using the volatility constraint, we obtain

$$\bar{\sigma}^2 = \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} = \frac{1}{\gamma^2} \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}. \quad (25)$$

Therefore,

$$\gamma = \frac{\sqrt{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}}{\bar{\sigma}}. \quad (26)$$

Substituting (26) into (24) gives

$$\mathbf{w} = \frac{\bar{\sigma}}{\sqrt{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} = \frac{\bar{\sigma}}{\lambda} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}. \quad (27)$$

The expected return of this portfolio is

$$\boldsymbol{\mu}'\mathbf{w} = \boldsymbol{\mu}' \left(\frac{\bar{\sigma}}{\lambda} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right) \quad (28)$$

$$= \frac{\bar{\sigma}}{\lambda} \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \quad (29)$$

$$= \frac{\bar{\sigma}}{\lambda} \lambda^2 \quad (30)$$

$$= \bar{\sigma} \lambda. \quad (31)$$

The volatility of this portfolio is

$$\sqrt{\mathbf{w}'\Sigma\mathbf{w}} = \sqrt{\frac{\bar{\sigma}^2}{\lambda^2}\boldsymbol{\mu}'\Sigma^{-1}\Sigma\Sigma^{-1}\boldsymbol{\mu}} \quad (32)$$

$$= \frac{\bar{\sigma}}{\lambda} \sqrt{\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}} \quad (33)$$

$$= \bar{\sigma}. \quad (34)$$

Therefore, the Sharpe ratio of the optimal portfolio is

$$\text{SR}(\mathbf{w}) = \frac{\boldsymbol{\mu}'\mathbf{w}}{\sqrt{\mathbf{w}'\Sigma\mathbf{w}}} \quad (35)$$

$$= \frac{\bar{\sigma}\lambda}{\bar{\sigma}} \quad (36)$$

$$= \lambda. \quad (37)$$

The same logic applies to the two-asset case. The maximum Sharpe ratio attainable by investing in the first two benchmark funds is given by

$$\lambda_2 = \sqrt{\boldsymbol{\mu}_2'\Sigma_2^{-1}\boldsymbol{\mu}_2}.$$

Since the first two assets are uncorrelated with the remaining $N - 2$ assets, we obtain that

$$\lambda^2 = \boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu} \quad (38)$$

$$= \begin{pmatrix} \boldsymbol{\mu}_2' & \boldsymbol{\mu}_{N-2}' \end{pmatrix} \begin{pmatrix} \Sigma_2^{-1} & 0 \\ 0 & \Sigma_{N-2}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_{N-2} \end{pmatrix} \quad (39)$$

$$= \boldsymbol{\mu}_2'\Sigma_2^{-1}\boldsymbol{\mu}_2 + \boldsymbol{\mu}_{N-2}'\Sigma_{N-2}^{-1}\boldsymbol{\mu}_{N-2} \quad (40)$$

$$= \lambda_2^2 + \boldsymbol{\mu}_{N-2}'\Sigma_{N-2}^{-1}\boldsymbol{\mu}_{N-2} \quad (41)$$

$$\geq \lambda_2^2. \quad (42)$$

Multiple Portfolio Optimization Problems Each portfolio $i \in \{R, D, S, A\}$ solves a distinct optimization problem in terms of: (i) the number of available asset classes, and (ii) the expected return vector perceived by investors, which may differ from the true vector $\boldsymbol{\mu}$. The reference portfolio R is the optimal portfolio of the first two benchmark funds, with volatility fixed at $\bar{\sigma}$, based on the vector of forecasted expected returns $\tilde{\boldsymbol{\mu}}_R$. The diversified portfolio D is the optimal portfolio of the N benchmark funds, with volatility fixed at $\bar{\sigma}$, based on the vector of forecasted expected returns $\tilde{\boldsymbol{\mu}}_D$. The strategic portfolio S is the optimal portfolio of the N benchmark funds, with volatility fixed at $\bar{\sigma}$, based on the vector of forecasted expected returns $\tilde{\boldsymbol{\mu}}_S$. Finally, the actual portfolio A has the same portfolio weights as the strategic portfolio but earns additional return from active management. We model the properties of active returns using the parameters α_A and $\sigma_{\epsilon, A}$, which respectively denote the active value-add and tracking error. The active return is uncorrelated with the returns of the N asset classes.

Appendix B.2 Transformed Economy and Portfolio Optimization

To model the distortion in return forecast specification and solve each portfolio, it is helpful to re-express the portfolio problem in an equivalent transformed economy by applying the linear transformation $\boldsymbol{\Sigma}^{-1/2}$ to the original asset returns. In this transformed economy, asset returns are uncorrelated and have unit variance, so the investment opportunity set is completely characterized by the expected return vector of the transformed assets. Portfolio optimization in the transformed economy yields the same moments for each portfolio $i \in \{R, D, S, A\}$, while providing a highly tractable framework for analyzing cases in which investors have inaccurate forecasts of expected returns. In this section, we introduce the transformed economy and explain why portfolio optimization in the transformed economy produces identical outcomes. We consider the case presented above where investors know the true expected return vector and optimize over the N benchmark funds.

Optimization in the Transformed Economy We define the vector of transformed returns as

$$\mathbf{r}_a = \Sigma^{-1/2} \mathbf{r}. \quad (43)$$

Then

$$\text{Var}(\mathbf{r}_a) = \Sigma^{-1/2} \Sigma \Sigma^{-1/2} = \mathbf{I}. \quad (44)$$

The vector of transformed asset weights is

$$\mathbf{w}_a = \Sigma^{1/2} \mathbf{w}, \quad (45)$$

so that

$$\mathbf{w} = \Sigma^{-1/2} \mathbf{w}_a. \quad (46)$$

The optimization (21) becomes

$$\max_{\mathbf{w}_a} \quad \mathbf{w}_a' \Sigma^{-1/2} \boldsymbol{\mu} \quad \text{s.t.} \quad \mathbf{w}_a' \mathbf{w}_a = \bar{\sigma}^2. \quad (47)$$

We define the transformed expected-return vector as

$$\boldsymbol{\mu}_a = \Sigma^{-1/2} \boldsymbol{\mu}. \quad (48)$$

The optimization (47) then becomes

$$\max_{\mathbf{w}_a} \quad \mathbf{w}_a' \boldsymbol{\mu}_a \quad \text{s.t.} \quad \mathbf{w}_a' \mathbf{w}_a = \bar{\sigma}^2. \quad (49)$$

We use this optimization procedure in the sections that follow.

Normalized Counterparts It will be useful to work with the normalized counterpart of the transformed expected-return vector $\boldsymbol{\mu}_a$, which we define as

$$\mathbf{u} = \boldsymbol{\mu}_a / \|\boldsymbol{\mu}_a\|, \quad (50)$$

where $\|\boldsymbol{\mu}_a\| = \sqrt{\boldsymbol{\mu}_a' \boldsymbol{\mu}_a} = \sqrt{\boldsymbol{\mu}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}} = \lambda$ and $\mathbf{u}' \mathbf{u} = 1$.

A similar transformation applies to the vector of expected returns of the first two asset classes. We define

$$\boldsymbol{\mu}_{a,2} = \boldsymbol{\Sigma}_2^{-1/2} \boldsymbol{\mu}_2. \quad (51)$$

We further define

$$\mathbf{u}_2 = \frac{\boldsymbol{\mu}_{a,2}}{\|\boldsymbol{\mu}_{a,2}\|}, \quad (52)$$

where $\|\boldsymbol{\mu}_{a,2}\| = \sqrt{\boldsymbol{\mu}_{a,2}' \boldsymbol{\mu}_{a,2}} = \sqrt{\boldsymbol{\mu}_2' \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2} = \lambda_2$. We then embed $\boldsymbol{\mu}_{a,2}$ and $\mathbf{u}_2$ into R^N by appending zeros for the additional $N - 2$ assets, so that the resulting vectors have the same dimensions as $\boldsymbol{\mu}_a$ and $\mathbf{u}$. Under the orthogonality assumption, we obtain that $\mathbf{u}' \mathbf{u}_2 = \frac{\lambda_2}{\lambda}$.

Appendix B.3 Portfolio Optimization without Estimation Errors

To provide intuition for the portfolio optimization when investors may have incorrect forecasts about expected returns, we begin with the simple case in which the expected return vector is estimated without error. Because $\tilde{\boldsymbol{\mu}}_D = \tilde{\boldsymbol{\mu}}_S = \boldsymbol{\mu}$, the portfolios D and S are identical. For convenience, we use the following notation. The vectors $\mathbf{u}_i$ and $\mathbf{u}_{a,i}$ correspond to

$$\mathbf{u}_i = \begin{cases} \mathbf{u}_2, & i = R, \\ \mathbf{u}, & i = D, S, A, \end{cases} \quad (53)$$

$$\mathbf{u}_{a,i} = \begin{cases} \mathbf{u}_{a,2}, & i = R, \\ \mathbf{u}_a, & i = D, S, A, \end{cases} \quad (54)$$

and the variable λ_i corresponds to

$$\lambda_i = \begin{cases} \lambda_2, & i = R, \\ \lambda, & i = D, S, A. \end{cases} \quad (55)$$

Optimal Portfolio Weights Portfolio i is constructed by choosing the transformed asset weights $\mathbf{w}_{a,i}$ to maximize the portfolio expected return for a target volatility $\bar{\sigma}$:

$$\max_{\mathbf{w}_{a,i}} \quad \mathbf{w}_{a,i}' \boldsymbol{\mu}_{a,i} \quad \text{s.t.} \quad \mathbf{w}_{a,i}' \mathbf{w}_{a,i} = \bar{\sigma}^2. \quad (56)$$

The Lagrangian is

$$\mathcal{L} = \mathbf{w}_{a,i}' \boldsymbol{\mu}_{a,i} - \frac{\gamma}{2} (\mathbf{w}_{a,i}' \mathbf{w}_{a,i} - \bar{\sigma}^2). \quad (57)$$

The first-order condition is

$$\boldsymbol{\mu}_{a,i} - \gamma \mathbf{w}_{a,i} = 0, \quad (58)$$

which implies

$$\mathbf{w}_{a,i} = \frac{1}{\gamma} \boldsymbol{\mu}_{a,i}. \quad (59)$$

Using the variance constraint, we get

$$\mathbf{w}_{a,i}' \mathbf{w}_{a,i} = \frac{1}{\gamma^2} \boldsymbol{\mu}_{a,i}' \boldsymbol{\mu}_{a,i} = \bar{\sigma}^2. \quad (60)$$

Since

$$\|\boldsymbol{\mu}_{a,i}\| = \sqrt{\boldsymbol{\mu}_{a,i}' \boldsymbol{\mu}_{a,i}} = \lambda_i, \quad (61)$$

we obtain

$$\gamma = \frac{\lambda_i}{\bar{\sigma}}. \quad (62)$$

Therefore, the optimal transformed-asset weights are

$$\mathbf{w}_{a,i}^* = \frac{\bar{\sigma}}{\lambda_i} \boldsymbol{\mu}_{a,i}. \quad (63)$$

Expected Returns Multiplying the vector of transformed weights $\mathbf{w}_{\mathbf{a},i}^*$ by the expected return vector of the transformed assets $\boldsymbol{\mu}_{\mathbf{a},i}$, we obtain the expected return of portfolio $i \in \{R, D, S\}$:

$$\mu_i = \begin{cases} \bar{\sigma}\lambda_2, & i = R \\ \bar{\sigma}\lambda, & i = D, S \end{cases} \quad (64)$$

For the expected return of the actual portfolio A , we add the active value-add, α_A , to the expected return of the strategic portfolio:

$$\mu_A = \mu_S + \alpha_A = \bar{\sigma}\lambda + \alpha_A.$$

Variances The variance of portfolio $i \in \{R, D, S\}$ is given by:

$$\sigma_i^2 = \mathbf{w}_{\mathbf{a},i}^{*'} \mathbf{w}_{\mathbf{a},i}^* = \bar{\sigma}^2. \quad (65)$$

For the variance of the actual portfolio, we add the tracking error variance, $\sigma_{\epsilon,A}^2$, to the variance of the strategic portfolio:

$$\sigma_A^2 = \sigma_S^2 + \sigma_{\epsilon,A}^2 = \bar{\sigma}^2 + \sigma_{\epsilon,A}^2.$$

Correlations For each pair of portfolios $(i, j) \in \{R, D, S\}$, the correlation is given by

$$\rho_{i,j} = \mathbf{w}_{\mathbf{a},i}^{*'} \mathbf{w}_{\mathbf{a},j}^* = \mathbf{u}_i' \mathbf{u}_j = \frac{\lambda_i}{\lambda_j} \quad (66)$$

We obtain the following expressions:

$$\rho_{i,j} = \begin{cases} \frac{\lambda_2}{\lambda} & \text{for } (i, j) = (R, D) \\ \frac{\lambda_2}{\lambda} & \text{for } (i, j) = (R, S) \\ 1 & \text{for } (i, j) = (D, S) \end{cases} \quad (67)$$

Because the return from active management can be written as $\alpha_A + \epsilon_A$ where ϵ_A is not correlated with the benchmark funds, we obtain the covariance between the returns of the strategic and actual portfolios as:

$$\sigma_{A,S} = Cov(r_A, r_S) = Cov(r_S + \alpha_A + \epsilon_A, r_S) = \sigma_S^2. \quad (68)$$

The correlation between both portfolios is therefore

$$\rho_{A,S} = \frac{\sigma_{A,S}}{\sigma_S \sigma_A} = \frac{\sigma_S^2}{\sigma_S \sigma_A} = \frac{\sigma_S}{\sigma_A} = \frac{\bar{\sigma}}{\sqrt{\bar{\sigma}^2 + \sigma_{\epsilon,A}^2}}. \quad (69)$$

Similarly, the correlation between the reference and actual portfolios is

$$\rho_{A,R} = \frac{\sigma_{A,R}}{\sigma_A \sigma_R} = \frac{\sigma_{S,R}}{\sigma_A \sigma_R} = \frac{\sigma_{S,R}}{\sigma_S \sigma_R} \frac{\sigma_S}{\sigma_A} = \rho_{S,R} \frac{\sigma_S}{\sigma_A} \quad (70)$$

$$= \left(\frac{\lambda_2}{\lambda} \right) \frac{\bar{\sigma}}{\sqrt{\bar{\sigma}^2 + \sigma_{\epsilon,A}^2}}. \quad (71)$$

Portfolio Optimization as a Source of Tracking Error Both portfolios A and D expand the investment opportunity set by adding assets that are uncorrelated with the existing assets. It may therefore seem surprising that the correlation between portfolios A and S differs from that between portfolios D and R . The key difference is that portfolio A retains the same benchmark weights as portfolio S , whereas portfolio D is fully reoptimized after the investment universe is expanded. As a result, the weights of portfolio D on equities and bonds generally differ from those of portfolio R , reducing the correlation between the two portfolios.

To illustrate the importance of this distinction, suppose instead that the entire portfolio is reoptimized after adding the active returns, and denote the resulting portfolio by A^* . Using the same arguments as above, its correlation with portfolio S is

$$\rho_{A^*,S} = \frac{\lambda}{\lambda_{N+1}}, \quad (72)$$

where $\lambda_{N+1} = \sqrt{\lambda^2 + IR^2}$ and $IR = \alpha_A/\sigma_{\epsilon,A}$ denotes the information ratio of the active strategy.

The correlation between A^* and S is lower than the correlation between A and S . To see why, recall that portfolio A is constructed by adding an orthogonal active return to portfolio S without changing its existing weights. Portfolio S is therefore the projection of A onto the space spanned by the original N assets. By the projection property, S has the highest correlation with A among all portfolios formed from these assets. In contrast, portfolio A^* is obtained by reoptimizing the entire portfolio after introducing the active investment opportunity. Portfolio S is therefore no longer the projection of A^* onto the original asset space, and its correlation with A^* must be lower. Hence,

$$\rho_{A^*,S} \leq \rho_{A,S}, \quad (73)$$

with equality only in the limiting case in which the introduction of the active opportunity does not alter the optimal weights on the original assets. This result highlights the importance of the projection property: adding an orthogonal return while keeping the benchmark weights unchanged mechanically produces a higher correlation with the benchmark than expanding the opportunity set and reoptimizing the entire portfolio.

Appendix B.4 Modelling Estimation Errors

We now assume that investors may have incorrect estimates about expected returns. Specifically, every portfolio $i \in \{R, D, S, A\}$ is based on a set of expected return forecasts for the N asset classes, which we denote by the vector $\tilde{\boldsymbol{\mu}}_i$. In this section, we specify $\tilde{\boldsymbol{\mu}}_i$.

We denote by $\mathbf{v}_i$ any $N \times 1$ unit vector orthogonal to $\mathbf{u}_i$:

$$\mathbf{u}_i' \mathbf{v}_i = 0, \quad \|\mathbf{v}_i\| = 1. \quad (74)$$

The expected return vector of the transformed assets associated with portfolio i is given

by:

$$\tilde{\boldsymbol{\mu}}_{a,i} = \lambda_i(\cos a_i \mathbf{u}_i + \sin a_i \mathbf{v}_i), \quad (75)$$

The vector $\mathbf{v}_i$ determines the direction of the forecast distortion, while the angle $a_i \in [0, \pi/2]$ measures its magnitude. When $a_i = 0$, the portfolio is constructed from the correct forecasts of expected returns, whereas larger values of a_i correspond to progressively greater distortions. In the original economy, the perceived expected return vector is therefore

$$\tilde{\boldsymbol{\mu}}_i = \boldsymbol{\Sigma}^{1/2} \tilde{\boldsymbol{\mu}}_{a,i} = \lambda_i \boldsymbol{\Sigma}^{1/2} (\cos a_i \mathbf{u}_i + \sin a_i \mathbf{v}_i). \quad (76)$$

We construct each portfolio from forecasts that differ in the direction of forecast distortion, $\mathbf{v}_i$, and the degree of distortion, a_i . The reference portfolio R is constructed from forecasts with direction vector $\mathbf{v}_R$ and mismeasurement level $\bar{a} \geq 0$. Similar to $\mathbf{u}_2$, we embed $\mathbf{v}_2$ into R^N by appending zeros for the additional $N - 2$ assets. The diversified portfolio D is constructed from forecasts with direction vector $\mathbf{v}_D$ and the same mismeasurement level, $\bar{a}$, reflecting that it is based on the same quality of information publicly available to individual investors. The strategic portfolio S is constructed from forecasts with direction vector $\mathbf{v}_S$ and a lower mismeasurement level $0 \leq a \leq \bar{a}$, capturing the fund's superior forecasting ability. The actual portfolio A has the same weights as the strategic portfolio S and therefore shares the same parameters $\mathbf{v}_S$ and a . We allow the error vectors $\mathbf{v}_R$, $\mathbf{v}_D$ and $\mathbf{v}_S$ to point in the same direction as measured by their scalar products denoted by $\rho_{R,D}^v$, $\rho_{D,S}^v$, and $\rho_{R,S}^v$.

We make two assumptions about the distortion vectors $\mathbf{v}_R$, $\mathbf{v}_D$ and $\mathbf{v}_S$.

Assumption 1: Each distortion vector is orthogonal to the true signal associated with its own opportunity set:

$$\mathbf{u}' \mathbf{v}_D = \mathbf{u}' \mathbf{v}_S = 0, \quad \mathbf{u}_2' \mathbf{v}_R = 0. \quad (77)$$

These conditions follow directly from the geometric decomposition of investors' perceived expected returns.

Assumption 2: The distortion vectors used to construct the diversified and strategic portfolios are orthogonal to the distortion vector used to construct the reference portfolio:

$$\mathbf{u}_2' \mathbf{v}_D = \mathbf{u}_2' \mathbf{v}_S = 0. \quad (78)$$

This isolates differences in investment opportunity sets without introducing differences in the component of belief distortions aligned with the reference portfolio.

All distortion vectors are normalized to unit length. We denote their pairwise scalar products by

$$\mathbf{v}_R' \mathbf{v}_D = \rho_{R,D}^v, \quad \mathbf{v}_R' \mathbf{v}_S = \rho_{R,S}^v, \quad \mathbf{v}_D' \mathbf{v}_S = \rho_{D,S}^v. \quad (79)$$

In the simulations, we set $\rho_{R,D}^v$ and $\rho_{R,S}^v$ to either 0.9 (baseline case) or 0.1, and $\rho_{D,S}^v$ to 0.9.

Under the block-diagonal covariance structure of $\mathbf{\Sigma}$, the first two components of the vector $\mathbf{u}$ are proportional to $\mathbf{u}_2$, with proportionality coefficient λ_2/λ . Hence,

$$\mathbf{u}' \mathbf{v}_R = \frac{\lambda_2}{\lambda} \mathbf{u}_2' \mathbf{v}_R = 0, \quad (80)$$

because $\mathbf{u}_2' \mathbf{v}_R = 0$ by construction. Therefore, $\mathbf{v}_R$ is automatically orthogonal to $\mathbf{u}$. Under the maintained assumptions, all three distortion vectors are consequently orthogonal to both $\mathbf{u}_2$ and $\mathbf{u}$.

We obtain the following set of forecasted expected returns of the transformed assets for portfolio $i \in \{R, D, S, A\}$:

$$\tilde{\mu}_{a,i} = \begin{cases} \lambda_2(\cos \bar{a} \mathbf{u}_2 + \sin \bar{a} \mathbf{v}_R), & i = R \\ \lambda(\cos \bar{a} \mathbf{u} + \sin \bar{a} \mathbf{v}_D), & i = D \\ \lambda(\cos a \mathbf{u} + \sin a \mathbf{v}_S), & i = S, A. \end{cases} \quad (81)$$

Our modeling of estimation errors has several appealing properties, which we describe below.

Property 1: Quality of Investment Opportunity Set is Unchanged The maximum Sharpe ratio perceived by investors $\tilde{\lambda}_i = \sqrt{\tilde{\boldsymbol{\mu}}_{a,i}' \tilde{\boldsymbol{\mu}}_{a,i}}$ is identical to the true maximum Sharpe ratio λ_i . That is, because $\mathbf{u}_i$ and $\mathbf{v}_i$ are orthonormal, we obtain

$$\tilde{\lambda}_i = \|\tilde{\boldsymbol{\mu}}_{a,i}\|^2 = \lambda_i^2(\cos^2(a_i) + \sin^2(a_i)) = \lambda_i^2. \quad (82)$$

Hence, investors are neither uniformly too optimistic nor uniformly too pessimistic about the market. Instead, they overestimate the attractiveness of some investment opportunities while underestimating that of others.

Property 2: Forecast Errors are Larger for Riskier Assets If we assume, for simplicity, that $\boldsymbol{\Sigma}$ is diagonal, the orthogonality condition between $\mathbf{u}_i$ and $\mathbf{v}_i$ can be written as

$$\sum_n \frac{\mu_n}{\sigma_n} v_n = 0, \quad (83)$$

where μ_n and σ_n is the return average and volatility of asset n and v_n is the corresponding distortion. All else equal, a higher volatility σ_n translates into a larger distortion v_n . Intuitively, high-volatility assets are more difficult to evaluate because their prices fluctuate more and therefore contain more noise. As a result, investors are more likely to misestimate their expected returns.

Appendix B.5 Portfolio Optimization with Estimation Errors

Building on (63), the optimal transformed-asset weights are

$$\tilde{\mathbf{w}}_{a,i}^* = \frac{\bar{\sigma}}{\lambda_i} \tilde{\boldsymbol{\mu}}_{a,i} = \bar{\sigma}(\cos a_i \mathbf{u}_i + \sin a_i \mathbf{v}_i). \quad (84)$$

Expected Returns Multiplying the portfolio weight vector $\tilde{\mathbf{w}}_{a,i}^*$ by the true expected return vector of the transformed assets $\boldsymbol{\mu}_{a,i}$, we obtain the expected return of portfolio

$i \in \{R, D, S\}$:

$$\mu_i = \tilde{\mathbf{w}}_{\mathbf{a},i}^*{}' \boldsymbol{\mu}_{\mathbf{a},i} \quad (85)$$

$$= \bar{\sigma}(\cos a_i \mathbf{u}_i + \sin a_i \mathbf{v}_i) \boldsymbol{\mu}_{\mathbf{a},i} \quad (86)$$

$$= \bar{\sigma} \lambda_i \cos a_i. \quad (87)$$

Applied to each portfolio $i \in \{R, D, S\}$, the expected return is

$$\mu_i = \begin{cases} \bar{\sigma} \lambda_2 \cos \bar{a}, & i = R \\ \bar{\sigma} \lambda \cos \bar{a}, & i = D \\ \bar{\sigma} \lambda \cos a, & i = S \end{cases} \quad (88)$$

For the expected return of the actual portfolio A , we add the active value-add to the expected return of the strategic portfolio:

$$\mu_A = \mu_S + \alpha_A = \bar{\sigma} \lambda \cos a + \alpha_A. \quad (89)$$

Variances The variance of portfolio $i \in \{R, D, S\}$ is given by:

$$\sigma_i^2 = \tilde{\mathbf{w}}_{\mathbf{a},i}^*{}' \tilde{\mathbf{w}}_{\mathbf{a},i}^* = \bar{\sigma}^2. \quad (90)$$

For the variance of the actual portfolio, we add the tracking error variance to the variance of the strategic portfolio:

$$\sigma_A^2 = \sigma_S^2 + \sigma_{\epsilon,A}^2 = \bar{\sigma}^2 + \sigma_{\epsilon,A}^2. \quad (91)$$

Correlations For each pair portfolios $(i, j) \in \{R, D, S\}$, the correlation is given by

$$\rho_{i,j} = \tilde{\mathbf{w}}_{\mathbf{a},i}^*{}' \tilde{\mathbf{w}}_{\mathbf{a},j}^* = \cos(a_i) \cos(a_j) \mathbf{u}_i{}' \mathbf{u}_j + \sin(a_i) \sin(a_j) \mathbf{v}_i{}' \mathbf{v}_j \quad (92)$$

We obtain the following expressions:

$$\rho_{i,j} = \begin{cases} \cos^2(\bar{a}) \frac{\lambda_2}{\lambda} + \sin^2(\bar{a}) \rho_{R,D}^v & \text{for } (i,j) = (R,D) \\ \cos(\bar{a}) \cos(a) \frac{\lambda_2}{\lambda} + \sin(\bar{a}) \sin(a) \rho_{R,S}^v & \text{for } (i,j) = (R,S) \\ \cos(\bar{a}) \cos(a) + \sin(\bar{a}) \sin(a) \rho_{D,S}^v & \text{for } (i,j) = (D,S) \end{cases} \quad (93)$$

The active portfolio A is based on the same parameters a and $\mathbf{v}_S$ as the strategic portfolio S but has a different volatility. Because the return from active management can be written as $\alpha_A + \epsilon_A$ where ϵ_A is not correlated with the benchmark funds, we obtain the correlation between both portfolios is therefore

$$\rho_{A,S} = \frac{\sigma_{A,S}}{\sigma_S \sigma_A} = \frac{\sigma_S^2}{\sigma_S \sigma_A} = \frac{\sigma_S}{\sigma_A} = \frac{\bar{\sigma}}{\sqrt{\bar{\sigma}^2 + \sigma_{\epsilon,A}^2}}. \quad (94)$$

Similarly, the correlation between the reference and actual portfolios is:

$$\rho_{A,R} = \frac{\sigma_{A,R}}{\sigma_A \sigma_R} = \frac{\sigma_{S,R}}{\sigma_A \sigma_R} = \frac{\sigma_{S,R}}{\sigma_S \sigma_R} \frac{\sigma_S}{\sigma_A} = \rho_{S,R} \frac{\sigma_S}{\sigma_A} \quad (95)$$

$$= \left(\cos(a) \cos(\bar{a}) \frac{\lambda_2}{\lambda} + \sin(a) \sin(\bar{a}) \rho_{R,S}^v \right) \frac{\bar{\sigma}}{\sqrt{\bar{\sigma}^2 + \sigma_{\epsilon,A}^2}}. \quad (96)$$

By a similar logic, the correlation between the correlation between the diversified and actual portfolios is:

$$\rho_{A,D} = \rho_{S,D} \frac{\sigma_S}{\sigma_A} = (\cos(\bar{a}) \cos(a) + \sin(\bar{a}) \sin(a) \rho_{D,S}^v) \frac{\bar{\sigma}}{\sqrt{\bar{\sigma}^2 + \sigma_{\epsilon,A}^2}}.$$

Appendix B.6 Computation of false-negative probability

Given the formulas for the moments of each portfolio pair $(i,j) \in \{R,D,S,A\}$ from the previous section, we can compute the false-negative probability $P(T)$ for the performance test comparing the average returns of two portfolios.

For a given portfolio pair (i, j) , we define the non-centrality parameter

$$z_{i,j} = \sqrt{T} \frac{\mu_i - \mu_j}{\sqrt{(\sigma_i - \sigma_j)^2 + 2\sigma_i\sigma_j(1 - \rho_{ij})}}. \quad (97)$$

We denote by $\bar{r}_i = \frac{1}{T} \sum_t r_i$ and $\bar{r}_j = \frac{1}{T} \sum_t r_j$ the average returns of portfolios i and j in the sample period. We can then measure the probability $P(T)$ that the observed spread between the two average returns in the sample period is negative under the hypothesis that the fund generates value-add along three possible dimensions: (i) diversification ($\lambda > \lambda_2$), (ii) strategic ($a < \bar{a}$), and (iii) active ($\alpha_A > 0$). If r_i and r_j are normally distributed, $P(T)$ is given by

$$P(T)_{i,j} = \text{prob}(\bar{r}_i - \bar{r}_j < 0 | \mu_i - \mu_j > 0) = \Phi(0, z_{i,j}, 1),$$

where $\Phi(0, z_{i,j}, 1)$ is the cumulative normal distribution function with a mean equal to $z_{i,j}$ and a unit variance.